MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
ZAPORIZHZHIA NATIONAL UNIVERSITY

O. H. CHEREP
Yu. V. KALIUZHNA
S. V. MARKOVA

FROM DEFENSE TO DEVELOPMENT: THE AI ECOSYSTEM FOR INFORMATION SECURITY AND ECONOMIC GROWTH

Monograph

IZDEVNIECĪBA
BALTIJA
PUBLISHING
2025

UDC 33(082)
Fr860

Reviewers:

OKHRIMENKO Ihor Vitaliiovych – Doctor of Economic Sciences, Professor, Rector of the Kyiv Cooperative Institute of Business and Law;

PULINA Tetiana Veniaminivna – Doctor of Economics, Professor, Head of the Department of Management and Administration at Zaporizhzhia Polytechnic National University;

SOBKO Olga Mykolaivna – Doctor of Economics, Professor, Head of the Department of Entrepreneurship and Trade at Western Ukrainian National University

Recommended by the decision of the Academic Council of Zaporizhzhia National University (protocol № 11 of 27.05.2025)

Fr860

From Defense to Development: the AI Ecosystem for Information Security and Economic Growth : monograph / edited by O. H. Cherep, Yu. V. Kaliuzhna, S. V. Markova. Riga, Latvia : Baltija Publishing, 2025. 202 p.

ISBN 978-9934-26-681-2

DOI https://doi.org/10.30525/978-9934-26-681-2

The monograph is devoted to a comprehensive analysis of artificial intelligence as a tool for the development of socio-economic security in the context of digitalization, labor market transformation, and the need to adapt to the challenges of wartime. It explores the definition, history of development, and main areas of application of artificial intelligence, as well as its potential in ensuring information resilience and effective management of big data. Special attention is paid to the experience of EU countries in countering disinformation and the possibilities of implementing similar solutions in the Ukrainian context.

The monograph is based on the results of research within the framework of the project of basic scientific research, applied scientific research, scientific and technical (experimental) developments on the topic № 2/25 "Artificial intelligence as a tool to counter disinformation during the war and post-war economic recovery in Ukraine" (state registration number 0125U000996) (01.01.2025–31.12.2027).

The monograph is intended for scholars, teachers, students of higher education institutions, graduate students, doctoral students, practitioners, representatives of state authorities and local self-government, business, university administrative staff, representatives of civil society, the public and all interested persons.

UDC 33(082)

ISBN 978-9934-26-681-2

МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Запорізький
Національний
Університет

О. Г. ЧЕРЕП
Ю. В. КАЛЮЖНА
С. В. МАРКОВА

ВІД ОБОРОНИ ДО РОЗВИТКУ: ЕКОСИСТЕМА ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ ТА ЕКОНОМІЧНОГО ЗРОСТАННЯ

Монографія

IZDEVNIECĪBA
BALTIJA
PUBLISHING
2025

УДК 004.8:[[001.102+004]-049.5+330.35]
Ч-46

Рецензенти:

Охріменко Ігор Віталійович – доктор економічних наук, професор, ректор Київського кооперативного інституту бізнесу і права;

Пуліна Тетяна Веніамінівна – доктор економіки, професор, завідувач кафедри менеджменту та адміністрації Запорізької політехнічної національної держави;

Собко Ольга Миколаївна – доктор економіки, професор, завідувач кафедри підприємництва та торгівлі Західноукраїнського національного університету

Рекомендовано до друку рішенням Наукової ради
Запорізького національного університету
(протокол № 11 від 27.05.2025 р.)

Череп О. Г.

Ч-46 Від оборони до розвитку: екосистема штучного інтелекту для інформаційної безпеки та економічного зростання : монографія / О. Г. Череп, Ю. В. Калюжна, С. В. Маркова. Рига, Латвія : Baltija Publishing, 2025. 202 с.

ISBN 978-9934-26-681-2

DOI https://doi.org/10.30525/978-9934-26-681-2

Монографія присвячена всебічному аналізу штучного інтелекту як інструменту розвитку соціально-економічної безпеки в контексті цифровізації, трансформації ринку праці та необхідності адаптації до викликів воєнного часу. Вона досліджує визначення, історію розробки та основні напрямки застосування штучного інтелекту, а також його потенціал для забезпечення інформаційної стійкості та ефективного управління великими даними. Особливу увагу приділяється досвіду країн ЄС у протидії дезінформації та можливостям впровадження подібних рішень в українському контексті.

Монографія базується на результатах досліджень у рамках проєкту фундаментальних наукових досліджень, прикладних наукових досліджень, науково-технічних (експериментальних) розробок на тему No2/25 «Штучний інтелект як інструмент протидії дезінформації під час війни та повоєнного економічного відновлення в Україні» (державний реєстраційний номер 0125U000996) (01.01.2025–31.12.2027).

Монографія призначена для науковців, викладачів, студентів вищих навчальних закладів, аспірантів, докторантів, практиків, представників державних органів влади та місцевого самоврядування, бізнесу, адміністративного персоналу університету, представників громадянського суспільства, громадськості та всіх зацікавлених осіб.

УДК 004.8:[[001.102+004]-049.5+330.35]

ISBN 978-9934-26-681-2

CONTENT

PREFACE

In the twenty-first century, artificial intelligence has moved beyond the status of a purely technological innovation and has become a strategic factor shaping national security, economic development, and global competitiveness. The rapid digitalization of societies, economies, and public institutions has created unprecedented opportunities for growth, efficiency, and innovation. At the same time, it has generated new vulnerabilities in the information space, including cyberattacks, disinformation campaigns, data breaches, and hybrid threats that challenge the stability of states and markets alike.

Traditionally, information security has been approached primarily from a defensive perspective – as a set of measures aimed at protection, risk mitigation, and damage control. However, this paradigm is no longer sufficient. Artificial intelligence enables a fundamental shift from reactive defense to proactive development, where information security becomes not only a protective shield, but also a catalyst for sustainable economic growth, innovation ecosystems, and resilience.

This monograph explores the concept of an AI-driven ecosystem in which information security, technological innovation, and economic development are deeply interconnected. It argues that AI technologies – such as machine learning, natural language processing, predictive analytics, and intelligent automation – can simultaneously strengthen information security and unlock new sources of value creation. When embedded in well-designed institutional, regulatory, and ethical frameworks, AI can enhance trust in digital environments, improve decision-making, foster entrepreneurship, and support long-term economic transformation.

Special attention is given to the role of AI in countering modern information threats, including cybercrime and large-scale

disinformation, while also supporting productivity growth, digital industries, and human capital development. The monograph emphasizes that the effectiveness of AI does not depend solely on algorithms and data, but on the broader ecosystem that includes governance structures, cross-sector collaboration, education, and societal readiness.

The purpose of this work is to provide a conceptual and analytical foundation for understanding how artificial intelligence can serve as a bridge between security and development. It is intended for researchers, policymakers, business leaders, and students interested in information security, digital transformation, and innovation-driven economic growth. By integrating insights from technology, economics, and security studies, this monograph seeks to contribute to a more holistic vision of AI as a tool not only for defense, but for building resilient, competitive, and future-oriented economies.

The monograph is based on the results of research within the framework of the project of basic scientific research, applied scientific research, scientific and technical (experimental) developments on the topic No. 2/25 “Artificial intelligence as a tool to counter disinformation during the war and post-war economic recovery in Ukraine” (state registration number 0125U000996) (01.01.2025–31.12.2027).

CHAPTER 1.
THEORY OF COMMUNICATIVE IMPACT AND DISINFORMATION

1.1. A MODEL OF COMMUNICATIVE IMPACT IN DISINFORMATION MESSAGES

Introduction. The modern information environment is characterized by the rapid spread of disinformation messages that appeal to emotions, provoke public anxiety, or distort perceptions of events. The main challenge lies in the difficulty of detecting such messages due to their semantic and emotional adaptability. Effective disinformation analysis requires a deep understanding of its linguistic structure, emotional tone, and pragmatic impact on the recipient's consciousness. In this context, artificial intelligence plays a particularly important role, as it can reveal hidden linguistic patterns and manipulative intentions.

Recent advancements in natural language processing (NLP) and machine learning have enabled artificial intelligence (AI) to perform increasingly complex tasks related to language understanding, including sentiment detection, discourse analysis, and text classification. These capabilities are particularly valuable in the context of disinformation, where the impact often hinges not on factual inaccuracy alone, but on emotional framing, subtle rhetorical manipulation, and the strategic deployment of linguistic elements to achieve perlocutionary effects.

Disinformation messages do not function merely as isolated texts; they operate as deliberate communicative acts that influence the cognitive, emotional, and behavioral states of target audiences. Understanding the structure and intent of these messages requires

more than surface-level content analysis. It involves unpacking the communicative functions embedded within the language-functions that are well-described by speech act theory. This theory offers a framework for analyzing not only what is said (locution), but why it is said (illocution), and with what effect it is said (perlocution).

In parallel, AI tools such as large language models (LLMs) provide scalable, data-driven mechanisms for identifying these communicative features across massive volumes of content. When combined, the theoretical precision of speech act analysis and the computational power of AI enable a more comprehensive and systematic approach to identifying disinformation. This synergy forms the foundation of the model proposed in this study.

Speech act theory (J. Austin, J. Searle) laid the foundation for understanding how language can not only transmit information but also influence behavior (Ostin J. L, 1986). In the context of disinformation, it is important to consider the three levels of a speech act: locution (the actual utterance), illocution (the author's intention), and perlocution (the real impact on the recipient). However, current disinformation studies still lack a systematic analysis of the emotional tone of such messages within these levels. Artificial intelligence tools, particularly NLP models, open new opportunities for detecting these types of influence.

Recent studies confirm that large language models (LLMs), including GPT-based architectures, are capable not only of analyzing but also of generating emotionally charged disinformation messages. It has been established, in particular, that the tone of prompts influences the level of manipulativeness in the models' responses, thereby enhancing the ability of AI to reproduce destructive communicative patterns (Vinay R., Oehmichen A., Agirre E., Davis B., 2024). Furthermore, a relevant research direction involves the analysis of disinformation networks and the identification of key actors in the dissemination of fake messages based on linguistic and behavioral patterns. This enables a comprehensive assessment

of disinformation impact not only at the content level but also within the structural dynamics of the information space (Smith S. T., Kao E. K., Mackin E. D., Shah D. C., Simek O., Rubin D. B., 2020).

In addition, recent reviews of the field highlight both theoretical and practical challenges in fake news detection, including the limitations of existing datasets, methodological constraints, and the necessity of forward-looking research agendas (Harris S., Hadi H. J., Ahmad N., Alshara M. A., 2024). At the same time, scholarly debate continues around the actual scope of generative AI's threat in the misinformation landscape, with some arguing that concerns about its impact may be overstated (Simon F. M., Altay S., & Mercier H., 2023), while others point to the heightened risks for journalism, public trust, and democratic discourse (Bell, E., 2023). Complementary research further demonstrates that AI systems can indeed produce persuasive propaganda, underscoring the dual-use nature of these technologies in the information domain (Goldstein J. A., Chao J., Grossman S., Stamos A., & Tomz M., 2023).

The aim of this study is to develop a methodological approach for identifying the emotional and manipulative structure of disinformation messages using speech act theory and artificial intelligence tools.

Presentation of the main research material. The methodological basis of this study is the integration of speech act theory with artificial intelligence (AI) tools for the purpose of modeling the communicative structure of disinformation messages. The approach is grounded in pragmatic linguistics, where disinformation is interpreted as a structured communicative act comprising three interrelated levels: locutionary, illocutionary, and perlocutionary. Each level is associated with distinct linguistic features and communicative functions.

The research design includes the following key stages:

1. Theoretical framework development: based on the works of J. Austin and J. Searle, a classification of speech acts is adapted for analyzing disinformation messages. Assertives, directives, commissives, and expressives are used as core illocutionary categories,

while verbal expressive devices (epithets, metaphors, anaphora, hyperbole) serve as indicators of perlocutionary effect.

2. Corpus formation: a set of real-world disinformation messages is compiled from open sources (e.g., social media, propaganda sites, messaging platforms). Each message is annotated manually and/or semi-automatically to identify pragmatic markers and emotional content.

3. Text preprocessing: NLP tools are used for tokenization, lemmatization, POS-tagging, and syntactic parsing. Stop words and noise elements are removed to ensure clean input for semantic analysis.

4. Locutionary analysis: this stage identifies the basic propositional structure of messages through the analysis of reference (nouns), predication (verbs), and grammatical cohesion. Semantic networks are used to extract central entities and actions.

5. Illocutionary analysis: modal verbs, discourse markers, syntactic structures, and illocutionary force indicators are used to classify the type of speech act (assertive, directive, etc.) using supervised machine learning models trained on annotated examples.

6. Perlocutionary analysis: the emotional and manipulative potential of verbal structures is assessed through sentiment analysis and emotion classification models (e.g., BERT, RoBERTa, or GPT-based). Lexical items are evaluated using emotion lexicons (e.g., NRC, LIWC) to determine their influence on recipient perception and possible behavioral effects.

7. Model construction: a layered analytical model is developed, mapping each component of the message to its corresponding speech act level. The model visualizes how linguistic elements contribute to the overall communicative and manipulative force of the disinformation.

8. Evaluation and validation: the model is tested on new message samples to assess its accuracy in detecting emotional tone, communicative intent, and manipulative constructs. Human-in-the-loop validation and cross-comparison with existing fact-checking datasets are used for evaluation.

The methodological novelty lies in the systematic coupling of speech act theory with computational linguistic analysis, enabling the automation of disinformation diagnostics not only at the lexical-semantic level, but also at the level of pragmatic intent. This methodology forms the foundation for building scalable tools in information security and cognitive resilience systems.

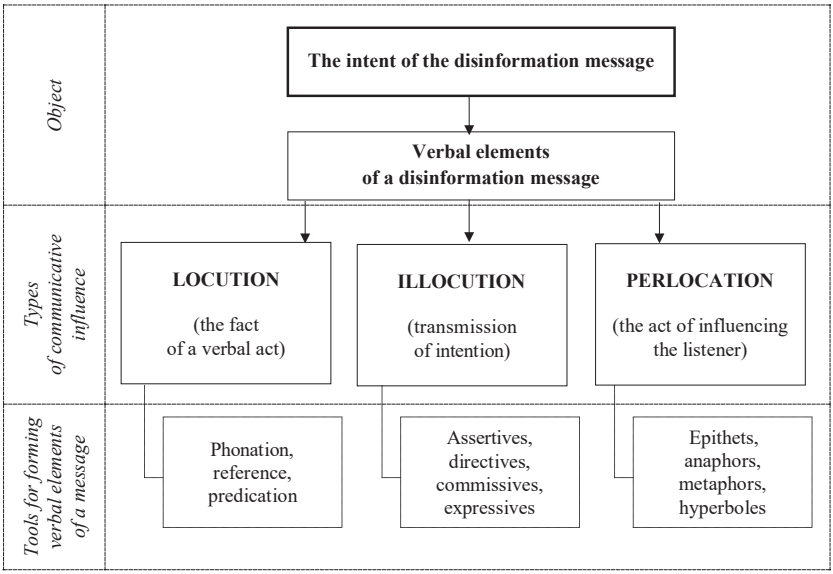
A disinformation message can be considered a communicative act that shapes the recipient's perceptions, emotions, and actions. The emotional tone of disinformation is a key indicator of its effectiveness. The formation of verbal elements is understood as a communicative act. The identification of the specifics of communicative intent in message formation is carried out from the perspective of speech act theory (J. Austin, 1986).

The authors propose to identify the locutionary, illocutionary, and perlocutionary components of the verbal elements of disinformation messages and to transpose the terminology traditionally applied to specific speech acts into the information environment (see Fig. 1.1, p. 8).

A communicative act is a rather complex phenomenon. According to speech act theory, there are three analytical levels or aspects of a speech act. First, a speech act can be viewed as the act of saying something. In this aspect, the speech act is treated as a locutionary act (from Latin *locution* – "speaking").

The locutionary act is itself a complex structure, as it includes sound production (phonation), word usage, syntactic arrangement, and the denotation of objects through language (reference), as well as the attribution of properties or relations to these objects (predication). Traditionally, linguistics has focused on the locutionary aspect of speech acts. However, the illocutionary and perlocutionary effects – especially in the context of disinformation tone analysis – have rarely been systematically explored, despite their potential to influence public consciousness both directly and indirectly.

Later, American philosopher (J. R. Searle, 1986), a student of J. Austin, in his article *A Taxonomy of Illocutionary Acts*, proposed



Intention (from Latin intentio – "aim, aspiration") refers to the direction of consciousness or thought toward a particular object.

Figure 1.1 – Methodological Approach to the Formation of Verbal Elements in Disinformation Messages (developed by the authors)

several categories of illocutionary acts that we also consider relevant for identifying the verbal elements in disinformation messages:

– Assertives – the speaker commits to the truth of the expressed proposition (e.g., to affirm, deny);

– Directives – the speaker attempts to get the listener to do something (e.g., request, command, permit, suggest);

– Commissives – the speaker commits to a future action (e.g., promise, swear, guarantee);

– Expressives – the speaker expresses a psychological state (e.g., thank, congratulate, apologize, sympathize).

Morphological units of emotive and evaluative vocabulary – such as epithets, anaphora, metaphors, and hyperbole – often manifest in adjectives, nouns, and verbs. These parts of speech most fully

convey the pragmatic orientation of a text (Grajs G. P., 1985). Their stylistic coloring helps emphasize important elements and draw attention to the desired emotional framing.

Nouns can construct conceptual images of the object, news, or message, enabling generalized associations for the recipient. Verbs, with their dynamic and expressive nature, serve to influence and prompt the recipient – thus enhancing the perlocutionary effect. Adjectives provide imagery, emotional tone, and stylistic richness, allowing the author to influence perception and imagination (Grishhenko A. I., 2007). Due to their high degree of subjectivity and emotionality, adjectives often function as tools of manipulation and are central to achieving disinformative impact.

From a linguistic perspective, effective communication is one in which the author's illocutionary intention results in a tangible perlocutionary effect.

The effectiveness of message perception depends on the composition and structuring of its components, the presence of a representational image, and an underlying idea or persuasive aim. The communicative function of informational appeals can be realized lexically (e.g., naming the recipient directly) and grammatically (e.g., using specific pronouns or imperative constructions).

Analyzing the internal structure of verbal elements and their communicative context is crucial. Two strategic errors may arise: insufficient attention to the recipient’s characteristics (resulting in communicative failure), or excessive adaptation (leading to the distortion of linguistic norms). Tailoring verbal elements to the intended audience enhances communicative impact. Interpretation of an information message often varies depending on the subject's personal perception.

From a communicative standpoint, verbal elements form components of a speech situation, where extralinguistic factors (topic, recipient, authorial intent) influence the selection of linguistic tools and, consequently, the genre and stylistic form of expression. Thus,

illocutionary force (authorial intention) manifests in news texts through specific codes – verbal, structural, and non-verbal.

Effective use of content-related components in verbal structures – especially those with strong locutionary and perlocutionary dimensions – enhances overall communication, allowing the message to reach its audience quickly and efficiently while shifting its tone from neutral information to targeted disinformation.

In emotionally charged disinformation contexts, Artificial Intelligence (AI) tools – particularly Natural Language Processing (NLP) methods – are vital analytical instruments. These technologies allow for automated analysis of textual content to detect emotional and manipulative potential.

One key direction is the identification of emotional tone. Pretrained language models such as BERT, RoBERTa, or GPT-like architectures can classify text into emotional categories including fear, anger, anxiety, outrage, compassion, etc. This enables the detection of messages with high emotional manipulation risk – especially those triggering mass anxiety or social tension.

The next step involves detecting manipulative linguistic patterns typical of disinformation content, including:

– excessive use of expressive vocabulary;
– rhetorical structures;
– stylistic exaggerations (hyperbole, metaphors, epithets);
– emotional appeals through contrast or personalization (e.g., “enemy – us,” “traitor – hero”);
– use of evaluative adjectives that impose subjective tone without evidence.

In addition, AI can help identify the illocutionary intentions of messages: assertions, calls to action, expressions of emotion, etc. This is achieved via analysis of syntactic structures, modal verbs, intention markers, and discourse formulas. Thus, AI is capable not only of detecting disinformation but also of interpreting the communicative intent of its author.

To ensure high accuracy, analytical systems combine:

– specialized emotion lexicons (e.g., NRC Emotion Lexicon, LIWC);

– corpora of annotated disinformation texts;

– audience response analysis modules (e.g., comments, engagement, sharing);

– in practice, such tools are implemented as:

• automated fact-checking instruments;

• content filtering or tagging modules on social media platforms;

• analytical support tools for information security professionals, journalists, and researchers.

Table 1.1 – Examples of Disinformation Message Analysis Based on Speech Act Theory and AI Tools

Fake news	AI tools used	Algorithm / module	AI-determined intent of a disinformation message	Possible impact on the audience
1	2	3	4	5
"Hundreds of people died after vaccination in India. This is the result of a global experiment on humanity!" (spread on social media, 2021)	NLP model analyzes semantics and tone: finds words with fear ("died", "experiment"), rhetorical structures	▲ Sentiment Analysis ▲ Named Entity Recognition ▲ Text Classification	▼ Emotion: fear, anger ▼ Illocution: assertive, expressive ▼ Manipulation: conspiracy theory, generalization	Fear of vaccination, undermining of trust in the healthcare system
"Ukrainian refugees receive more aid than locals!" (propaganda sites, EU, 2022)	AI performs lexical and emotional analysis: highlights comparisons,	▲ Emotion Detection ▲ Text Framing Detection	▼ Emotion: envy, indignation ▼ Illocution: directive ▼ Manipulation: social opposition	Provocation of hatred, social tension between refugees and locals

End of Table 1.1

1	2	3	4	5
	emotionally colored adjectives (“more”, “dishonest”), emphasis on conflict between groups	▲ Topic Modeling		
“The US army created the virus in a laboratory – documents confirm this!” (Russian disinformation, 2020–2021)	Model looks for fake structures: no evidence, references to “documents”, exaggeration, “conspiracy” semantics, classifier of typical fakes is applied	▲ Fake News Classifier ▲ Evidence Checker ▲ Semantic Similarity Search	▼ Emotion: suspicion, anger ▼ Illocution: assertive ▼ Manipulation: pseudo-facts, conspiracy, fear	Distrust of international partners, change of geopolitical loyalties
“Europe is on the verge of collapse: Ukrainians have destroyed the economy!” (anonymous Telegram channels, 2022–2023)	AI identifies exaggeration, hostile rhetoric, emotional tension; detects anonymous source as a marker of unreliability	▲ Hyperbole Detector ▲ Hate Speech Detection ▲ Source Credibility Scoring	▼ Emotion: fear, anger ▼ Illocution: directive, expressive ▼ Manipulation: hyperbole, xenophobia	Increasing anti-Ukrainian sentiment, radicalization of part of the audience

Source: compiled by the authors

Algorithm for Disinformation Message Analysis Based on Speech Act Theory and AI Tools

1. Preprocessing the text: tokenization, noise and tag removal, word normalization.

2. Determining the intent of the disinformation message; identifying types of communicative impact, key lexemes, and frames (e.g., triggers like fear, conflict, enemy, conspiracy, call to action).

3. Identifying tools for verbal element formation:

– Locution: phonation, reference, predication;

– Illocution: assertives, directives, commissives, expressives;

– Perlocution: epithets, anaphora, metaphors, hyperbole.

4. Assessing reliability: source verification, reference checking, fact-checking.

5. Estimating the potential impact: assessment of emotional, social, and political harm.

Conclusions. The study has established that disinformation messages possess a complex communicative structure based on the realization of the locutionary, illocutionary, and perlocutionary levels of speech acts. The primary tool of influence is the verbal content of the message – lexical units with emotional and evaluative connotations that amplify the psychological effect of the information and shape how it is perceived by the audience.

Artificial intelligence – particularly natural language processing (NLP) technologies – provides effective means for detecting:

– the emotional tone of messages (e.g., fear, anger, anxiety, compassion);

– types of illocutionary intentions (e.g., directives, assertions, emotional expressions);

– key semantic constructions that constitute manipulative content.

This approach enables a deeper analysis of disinformation texts not only at the content level but also through the identification of their pragmatic goals and potential influence on mass consciousness.

The integration of speech act theory with AI opens new avenues for the automated detection and mitigation of harmful informational influences, thereby enhancing the resilience of the information environment.

Thus, the use of artificial intelligence in disinformation analysis represents a promising direction that allows for the integration of linguistic, psychological, and technological dimensions in the development of systems designed to counter emotionally charged manipulative influence.

References

1. Ostin, J. L. (1986). Slovo yak diya [Word as action]. *Nove v zarubizhnij lingvisticzi. Teoriya movnikh aktiv – New in foreign linguistics. Theory of Speech Acts*, 17, 22–129 [in Ukrainian].

2. Serl, J. (1986). Shho take movnij akt [What is a speech act]. *Nove v zarubizhnij lingvisticzi: teoriya movnikh aktiv – New in foreign linguistics: theory of Speech Acts*, 17, 151–169 [in Ukrainian].

3. Grajs, G. P. (1985). Logika i movne spilkuvannya [Logic and Speech Communication]. *Nove v zarubizhnij lingvisticzi: lingvistichna pragmatika – New in foreign linguistics: linguistic pragmatics*, 16, 217–237 [in Ukrainian].

4. Grishhenko, A. I. (2007). Dzherela viniknennya ekspresivnikh etnonimiv (etnofolizm) v suchasnij rosijskij i anglijskij movakh: etimologichnij, motivaczijnij i derivaczijnij aspekti [Sources of the emergence of expressive ethnonyms (ethnofolisms) in modern Russian and English: etymological, motivational and derivational aspects]. *Materiali Mizhnarodnoyi konferencziyi pamyati L. V. Nikolenko ta Y. P. Solodub "Aktivni proczesi v suchasnij leksiczi i frazeologiyi" – Materials of the International Conference in Memory of L. V. Nikolenko and Y. P. Soloduba "Active processes in modern vocabulary and phraseology"* (pp. 40–52). Yaroslavl: TOV "Remder" [in Ukrainian].

5. Borisova, E. G. (2001). Perlokutivnij lingvistika ta yiyi vikladannya studentam-filologam [Perlocutionary Linguistics and its Teaching to Philology Students]. *Visnik Moskovskogo universitetu – Bulletin of Moscow University*, 1, 115–133 [in Ukrainian].

6. Cherep, O., Kaliuzhna, Y., & Markova, S. (2025). Model of communicative impact in disinformation messages based on speech act theory and artificial intelligence tools. *Baltic Journal of Economic Studies*, *11*(4), 410–415. https://doi.org/10.30525/2256-0742/2025-11-4-410-415

7. Vinay, R., Oehmichen, A., Agirre, E., Davis, B. (2024). Emotional Manipulation Through Prompt Engineering Amplifies Disinformation Generation in AI Large Language Models. *ArXiv preprint arXiv:2405.15923*. URL: https://arxiv.org/abs/2403.03550

8. Smith, S. T., Kao, E. K., Mackin, E. D., Shah, D. C., Simek, O., Rubin, D. B. (2020). Automatic Detection of Influential Actors in Disinformation Networks. *ArXiv preprint arXiv:2010.11920*. URL: https://arxiv.org/abs/2005.10879

9. Harris, S., Hadi, H. J., Ahmad, N., Alshara, M. A. Fake News Detection Revisited: An Extensive Review of Theoretical Frameworks, Dataset Assessments, Model Constraints, and Forward-Looking Research Agendas. *Technologies* 2024, 12, 222. https://doi.org/10.3390/technologies12110222

10. Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-127

11. Bell, E. (2023, March 3). Fake news, ChatGPT, truth, journalism, disinformation. *The Guardian*. https://www.theguardian.com/commentisfree/2023/mar/03/fake-news-chatgpt-truth-journalism-disinformation

12. Goldstein, J. A., Chao, J., & Grossman, S., Stamos, A. & Tomz, M. (2023). Can AI write persuasive propaganda? SocArXiv. https://doi.org/10.31235/osf.io/fp87b

1.2. SPEECH ACTS AND NARRATIVE STRUCTURES OF DISINFORMATION

Introduction. In an era defined by rapid digital communication and widespread access to information, the phenomenon of disinformation has emerged as a critical challenge to public discourse, democratic processes, and social cohesion. Disinformation, often deliberately crafted to mislead or manipulate, operates not merely through the content of messages but through the ways in which these messages are structured and delivered. Understanding disinformation requires a dual focus: the linguistic strategies employed by communicators and the narrative frameworks that make such content persuasive or believable.

Speech act theory provides a valuable lens for analyzing disinformation, as it enables scholars to examine how language performs actions beyond conveying information – such as persuading, threatening, promising, or deceiving. By identifying the illocutionary and perlocutionary forces embedded in misleading messages, researchers can uncover the pragmatic mechanisms through which disinformation achieves its effects. Simultaneously, narrative structures – the plots, characters, and causal relationships that organize information – play a crucial role in shaping audience interpretation. Disinformation often relies on coherent, emotionally resonant narratives that exploit cognitive biases, preexisting beliefs, and social identities.

This article explores the intersection of speech acts and narrative structures in the dissemination of disinformation. By integrating insights from linguistics, communication studies, and cognitive psychology, it aims to illuminate how deceptive messages are crafted, how they gain traction, and why they can be remarkably resistant to correction. Understanding these dynamics is essential for developing effective strategies to detect, counter, and ultimately mitigate the societal impacts of disinformation.

Presentation of the main research material. The digital age has dramatically transformed the landscape of information dissemination, enabling rapid, large-scale communication through social media, messaging platforms, and online news outlets. This transformation, however, has also facilitated the spread of disinformation, defined as deliberately false or misleading information intended to deceive.

Disinformation is not merely a collection of false statements; it is a carefully crafted communicative act. Its effectiveness depends not only on the factual content but also on the way it is presented, structured, and embedded within broader narratives that resonate with audiences.

Traditional analyses of disinformation often focus on fact-checking and content verification. While these approaches are valuable, they overlook the linguistic and narrative mechanisms that make false information persuasive and resistant to correction. Understanding these mechanisms requires an integrated approach, combining insights from pragmatics, narrative theory, and social psychology.

Speech act theory provides a framework for analyzing the pragmatic dimension of disinformation. According to Austin (1962) and Searle (1969), utterances perform actions beyond merely conveying information. In the context of disinformation, speech acts can function to persuade, alarm, delegitimize, or mobilize audiences.

Illocutionary acts – what the speaker intends to do by saying something – are particularly relevant to disinformation. For example, a false news report might perform a warning, accusation, or promise, thereby generating specific cognitive and emotional responses regardless of the statement's factual accuracy.

Perlocutionary acts – how audiences are affected by a statement – highlight the manipulative potential of disinformation. A message's success depends on its ability to provoke fear, reinforce biases, or incite action, demonstrating that disinformation operates on both cognitive and behavioral levels.

Beyond individual speech acts, narrative structures play a critical role in shaping the perception and impact of disinformation. Narratives provide

coherence, causal logic, and emotional engagement, enabling audiences to integrate isolated false claims into a seemingly credible story.

Common narrative strategies in disinformation include constructing clear villains and heroes, establishing cause-and-effect sequences, and appealing to preexisting cultural or political schemas. These structures exploit cognitive shortcuts, making messages more memorable and persuasive.

Disinformation often employs repetition, framing, and emotional language to reinforce narratives. Repetition increases familiarity, framing guides interpretation, and emotionally charged language triggers automatic responses, creating a feedback loop that strengthens belief in the narrative.

The interplay between speech acts and narrative structures is central to understanding disinformation. While speech acts perform the immediate communicative function, narrative structures provide the scaffolding that maintains and amplifies the message over time.

Analysis of disinformation requires examining how specific speech acts are embedded within broader narratives. For example, a statement accusing a public figure of corruption may function as an assertive illocutionary act within a narrative portraying widespread institutional decay.

Cognitive and social factors further enhance the effectiveness of these structures. Confirmation bias, motivated reasoning, and in-group identification make audiences more receptive to narratives that align with their existing beliefs, regardless of the veracity of individual claims.

Understanding the structural patterns of disinformation has practical implications for countermeasures. By identifying recurring speech acts and narrative frameworks, researchers and policymakers can design targeted interventions, such as inoculation strategies, media literacy programs, and automated detection systems.

This article aims to provide a systematic examination of disinformation through the dual lens of speech acts and narrative

structures. It seeks to map the mechanisms by which false information is constructed, communicated, and internalized, offering both theoretical insights and practical applications.

Ultimately, addressing the challenges of disinformation requires moving beyond surface-level fact-checking. A nuanced understanding of the linguistic and narrative architecture of deceptive messages is essential for developing effective strategies to mitigate their social, political, and cognitive impacts.

Structure of Disinformation

1. Headline / Hook.

Function: Captures attention, evokes an emotional reaction, sets the tone for the story.

Features:

– Vivid, emotionally charged language.
– Often formulated as a statement or warning.
– May contain implicit assertions ("you need to know…", "things are not as they seem").

Speech act: Assertive, Directive (urging attention or action).

2. Problem Introduction (Problem Setup).

Function: Presents context, frames a "crisis" or "threat."

Features:

– Mixes real facts with distorted or false information.
– Creates a sense of urgency or danger.
– Speech act: Commissive (promise to reveal the truth), Directive (prompting attention).

3. Main Characters.

Function: Establishes "heroes" and "villains," simplifies moral evaluation of events.

Features:

– Victims and enemies are often highly polarized.
– Anonymity or vagueness can be used to increase distrust of official sources.
– Narrative role: Hero / Villain.

4. Main Plot (Causal Narrative / Storyline).

Function: Creates cause-and-effect logic explaining events.

Features:

– Explains "why" and "how" events occurred.

– May rely on conspiratorial elements or distorted data.

– Speech act: Assertive (stating facts), Directive (leading the audience to a conclusion).

5. Emotional Hooks (Emotional Appeals).

Function: Enhances engagement, mobilizes emotions (fear, anger, anxiety).

Features:

– Uses vivid adjectives, dramatization, and personal stories.

– Amplifies feelings of injustice or threat.

– Speech act: Expressive (expressing emotion), Directive (prompting reaction).

6. Repetition of Key Ideas (Repetition / Framing).

Function: Strengthens memorability and creates cognitive schemas.

Features:

– Repetition of key statements and messages.

– Often paired with visual elements (memes, charts).

– Narrative role: Reinforcement.

7. Use of Authorities and "Evidence" (Evidence / Testimonials).

Function: Creates an illusion of credibility.

Features:

– Combines truthful information with fabricated content.

– Quotes from "experts," "leaked data," or references to questionable sources.

– Speech act: Assertive (asserting facts).

8. Manipulative Questions and Implications (Leading Questions / Implicature).

Function: Guides the audience to draw conclusions without explicit statements.

Features:

– "What if the truth is not what they say?"

– Encourages doubt toward official sources.

– Speech act: Directive (prompting thought), Assertive (implying a fact).

9. Climax (Peak Emotional Impact).

Function: Maximizes emotional effect and persuasion.

Features:

– Presents the most shocking or convincing part of the story.

– Often accompanied by visual or sensational elements.

– Narrative role: Turning point / Highlight.

10. Conclusion / Moral (Conclusion / Call to Action).

Function: Wraps up the narrative and directs the audience to action or belief.

Features:

– Encourages sharing, distrust of official sources, or support for a certain position.

– Speech act: Directive (urging action), Commissive (promising "truth" if recommendations are followed).

11. Use of Visual and Multimedia Elements.

Function: Reinforces the narrative, creates emotional and cognitive anchors.

Features:

– Memes, videos, graphics, fake screenshots.

– Strengthens the illusion of credibility through "evidence."

12. Cross-Platform Repetition (Multi-platform Amplification).

Function: Creates the illusion of consensus and widespread acceptance.

Features:

– Same narrative repeated across social media, forums, messaging apps.

– Enhances the effect of social proof.

13. References to "Anti-Sources".

Function: Undermines trust in verified information.

Features:

– Rhetoric such as “they are hiding the truth” or “official media is lying.”

– Creates cognitive dissonance in the audience.

14. Reinforcement through Archetypes.

Function: Builds a recognizable pattern for future messages.

Features:

– Standardized images of “victims” and “enemies.”

– Narrative easily adapted to new events.

15. Interactive Audience Engagement.

Function: Encourages active participation (likes, shares, comments).

Features:

– Uses questions, polls, challenges.

– Strengthens social proof and engagement loops.

Types of Disinformation

1. Fabricated Content.

Definition: Completely false information created with the intent to deceive.

Features:

– No factual basis; entirely invented stories or data.

– Often highly sensational to attract attention.

– Example: Fake news about a celebrity death that never occurred.

– Speech act: Assertive (stating a fact), Directive (prompting belief or action).

2. Manipulated Content.

Definition: Genuine information or images altered to mislead.

Features:

– Edited photos, videos, or audio to change meaning.

– Misrepresentation of facts to fit a false narrative.

Example: Deepfake video of a politician making a controversial statement.

Speech act: Assertive (implying truth), Expressive (evoking emotion).

3. Imposter Content.

– Definition: Information presented as coming from a credible source but is fake.

Features:

– Fake social media accounts, spoofed websites, or forged press releases.

– Exploits trust in authoritative sources.

Example: Fake news article mimicking the layout of a reputable newspaper.

Speech act: Assertive (pretending to state facts), Directive (guiding perception).

4. Misleading Content.

Definition: Information that is technically true but presented in a misleading context.

Features:

– Cherry-picked facts, selective reporting, or deceptive framing.

– Can create false impressions or conclusions.

Example: Reporting statistics without context to exaggerate trends.

Speech act: Assertive (partial truth), Directive (steering interpretation).

5. False Context.

Definition: Genuine content placed in a false context to mislead.

Features:

– Real images or quotes used to support unrelated claims.

– Often combines visuals with a fabricated narrative.

Example: A photo from a past event presented as happening now.

Speech act: Assertive (misrepresenting context), Directive (shaping understanding).

6. Satire or Parody (misinterpreted as truth).

Definition: Content intended as humor or satire but mistaken as factual.

Features:

– Exaggeration and irony used for critique or entertainment.

– Risk of spreading when audience misinterprets intent.

Example: Satirical news site story shared as real news.

Speech act: Expressive (humor, critique), Assertive (misunderstood as fact).

7. Conspiracy Theories.

Definition: Complex narratives claiming hidden, malevolent schemes.

Features:

– Often resistant to evidence, relies on anecdotal or circumstantial "proof."

– Provides a causal explanation for events, appealing to cognitive biases.

Example: Claims that global events are secretly orchestrated by a shadow organization.

Speech act: Assertive (alleged fact), Directive (mobilizing belief).

8. Clickbait / Sensationalism.

Definition: Misleading headlines or teasers designed to attract clicks.

Features:

– Often exaggerates or distorts content to maximize engagement.

– May link to partially true or unrelated content.

Example: "You won't believe what this politician said!" linking to unrelated statements.

Speech act: Directive (prompting interaction), Assertive (creating false expectation).

9. Propaganda.

Definition: Information disseminated to advance a political or ideological agenda.

Features:

– Appeals to emotions, identity, or loyalty rather than facts.

– Often combines multiple disinformation types.

Example: Government-sponsored narratives portraying opponents negatively.

Speech act: Assertive (claims), Directive (mobilizing support), Expressive (evoking emotion).

Table 1.2 – Types of Disinformation

Type	Definition	Features	Example	Speech Acts	Narrative Techniques
1	2	3	4	5	6
Fabricated Content	Completely false information created to deceive	No factual basis, highly sensational	Fake news about a celebrity death that never occurred	Assertive, Directive	Shock value, exaggerated plot
Manipulated Content	Genuine information or media altered to mislead	Edited photos / videos, distorted meaning	Deepfake video of a politician making a controversial statement	Assertive, Expressive	Visual manipulation, selective framing
Imposter Content	Information mimicking a credible source	Fake social media accounts, spoofed websites	Fake article mimicking a reputable newspaper	Assertive, Directive	Authority mimicry, trust exploitation
Misleading Content	Technically true information presented misleadingly	Cherry-picked facts, selective reporting	Statistics presented out of context to exaggerate trends	Assertive, Directive	Deceptive framing, selective emphasis
False Context	Real content placed in a false context	Misrepresented images, quotes	Photo from past event presented as current	Assertive, Directive	Context distortion, re-framing
Satire / Parody	Humorous or ironic content misinterpreted as true	Exaggeration, irony	Satirical news story shared as factual	Expressive, Assertive	Irony, exaggeration

End of Table 1.2

1	2	3	4	5	6
Conspiracy Theories	Complex narratives claiming hidden schemes	Resistant to evidence, anecdotal proofs	Claims of a shadow organization orchestrating events	Assertive, Directive	Causal narrative, villain-hero archetypes
Clickbait / Sensationalism	Misleading headlines designed to attract clicks	Exaggerated or distorted headlines	"You won't believe what this politician said!" linking to unrelated content	Directive, Assertive	Curiosity hooks, suspense, exaggeration
Propaganda	Information advancing political / ideological agenda	Emotional appeals, identity framing	Government narratives portraying opponents negatively	Assertive, Directive, Expressive	Emotional narrative, hero-villain framing, repetition

Source: compiled by the authors

Disinformation is not simply "falsehood"; it is a carefully designed communicative act that combines speech acts and narrative strategies to influence an audience. Looking at the table of types, it becomes clear that each form of disinformation relies on a specific combination of these tools.

First, the role of speech acts.

For instance, fabricated content and imposter content primarily employ assertive acts, presenting false information as fact. At the same time, directive acts often encourage the audience to take action: sharing, commenting, or distrusting official sources.

Expressive acts are more common in manipulated content, satire/parody, and propaganda, where the emotional impact is as important as the informational content.

This dual mechanism explains why fact-checking alone is often insufficient: the audience is already emotionally invested in the narrative, which makes cognitive corrections less effective.

Second, the role of narrative structures.

Nearly all types of disinformation rely on well-structured narratives. Even clickbait contains a "hook," emotional climax, and an implicit call to action.

Conspiracy theories and propaganda represent full-fledged narratives with heroes, villains, cause-effect relationships, and morals. These structures create a coherent cognitive world in which the audience accepts the story as a whole, not just individual claims.

Third, psychological and social factors.

Mechanisms such as confirmation bias, social proof, and emotional contagion amplify the impact of disinformation. For example, repetition and multi-platform dissemination create the illusion of consensus, particularly important for imposter content and false context.

Emotional triggers make even partially false information "plausible" because the human brain responds more strongly to emotional salience than to logical verification.

Fourth, strategic flexibility.

Types of disinformation rarely occur in isolation. A single message may combine fabricated content + propaganda + false context to enhance credibility. This multi-layered structure makes disinformation particularly difficult to detect and counter.

Conclusions. Disinformation is a sophisticated communicative phenomenon that operates not only through false content but through the strategic combination of speech acts and narrative structures. Its effectiveness depends on both cognitive persuasion and emotional engagement.

Speech acts – including assertive, directive, and expressive forms – serve as the primary mechanisms by which disinformation asserts claims, mobilizes audiences, and evokes emotional responses. Recognizing these acts is essential for understanding how deceptive messages influence beliefs and behavior.

Narrative structures provide coherence and context, transforming isolated false statements into compelling stories. Elements such as heroes and villains, cause-effect chains, and emotionally charged plots enhance memorability, believability, and audience investment.

Different types of disinformation – fabricated content, manipulated content, imposter content, misleading content, false context, satire, conspiracy theories, clickbait, and propaganda – employ distinct combinations of speech acts and narrative techniques. Some types, like propaganda or conspiracy theories, rely on multi-layered narratives to sustain long-term influence.

Emotional manipulation and cognitive biases, such as confirmation bias and the reliance on social proof, significantly amplify the impact of disinformation. Repetition, cross-platform dissemination, and visual cues strengthen the perceived credibility of false narratives.

The interplay between speech acts and narrative structures explains why disinformation is often resistant to fact-checking alone. Emotional engagement and narrative coherence frequently override

logical evaluation, making audiences more likely to internalize misleading messages.

Effective countermeasures must go beyond content verification. Strategies should include:

– mapping recurring narrative patterns in disinformation campaigns;

– identifying speech acts designed to manipulate emotions or prompt specific actions;

– educating audiences through media literacy and inoculation techniques that reduce susceptibility to cognitive biases.

Ultimately, understanding the architecture of disinformation – how language performs action and narratives shape perception – is crucial for mitigating its societal, political, and psychological impacts. Interdisciplinary approaches combining linguistics, communication studies, and cognitive psychology offer the most promising path forward.

References

1. Maurer, Peter (2019). "In the grip of politics? How political journalists in France and Germany perceive political influence on their work". Journalism, v. 20, n. 9. https://doi.org/10.1177/1464884917707139

2. Liotti, Jorge (2014). "The complex relationship between the media and the political system in Argentina: From co-option to polarization". In: Guerrero, Manuel-Alejandro; Márquez-Ramírez, Mireya (eds.). Media systems and communication policies in Latin America. London: Palgrave Macmillan UK. Pp. 100–121. ISBN: 9781349488476. https://doi.org/10.1057/9781137409058_6

3. Hallin, Daniel C., Mancini, Paolo (2012). "Comparing media systems" between Eastern and Western Europe". In: Gross, Peter; Jakubowicz, Karol (eds.). Media transformations in the post-communist world: Eastern Europe's tortured path to change. Lexington Books, pp. 15–32. ISBN: 9780739174944.

4. Del-Vicario, Michela, Quattrociocchi, Walter, Scala, Antonio, Zollo, Fabiana (2019). "Polarization and fake news: Early warning of potential

misinformation targets". ACM transactions on the Web, v. 13, n. 2. https://doi.org/10.1145/3316809

5. Kane, E. (2003). Continuing Dangers of Disinformation in Corporate Accounting Reports. nber.org. Working Paper, 9634. URL: https://www.nber.org/papers/w9634.pdf

6. Hou Zhiuan, Du Fanxing, Jiang Hao, Zhou Xinyu, Lin Lessa, Assessment, T., & Commission, N. H. (2020). Assessment of public attention, risk perception, emotional and behavioural responses to the COVID-19 outbreak: social media surveillance in China. medRxiv 2020.03.14.20035956. https://doi.org/10.1101/2020.03.14.20035956

7. Scheufele, A. Dietram, Krause M. Nicole (2019). Science audiences, misinformation, and fake news. Proceedings of the National Academy of Sciences of the United States of America 116(16), 7662–9. Accessed November 20, 2020. https://doi.org/10.1073/pnas.1805871115

8. Nougayrede, Natalie (2018). In this age of propaganda, we must defend ourselves. Here's how. The Guardian (31/01/18. https://www.theguardian.com/commentisfree/2018/jan/31/propaganda-defend-russia-technology

CHAPTER 2.
ECONOMIC CONSEQUENCES OF DISINFORMATION AND INDICATORS OF INFORMATION RESILIENCE

2.1. MODELS OF DISINFORMATION'S IMPACT ON TRUST, MARKETS, AND MACRO INDICATORS

Introduction. In recent years, disinformation has emerged as a critical factor shaping the global socio-economic environment. The rapid expansion of digital communication channels, social media platforms, and algorithm-driven content distribution has dramatically increased both the speed and scale at which false, misleading, or strategically distorted information can spread. As a result, disinformation is no longer confined to political discourse or isolated social groups, but increasingly penetrates economic decision-making, financial markets, and macroeconomic dynamics. This transformation calls for a systematic analytical framework capable of capturing the complex interactions between information integrity and economic outcomes.

Trust plays a foundational role in the functioning of modern economies. Confidence in institutions, markets, data providers, and policy authorities underpins expectations, coordination, and long-term investment decisions. Disinformation can erode this trust by undermining the credibility of central banks, governments, financial intermediaries, and statistical agencies. When economic agents begin to question the reliability of information, uncertainty rises, risk perceptions change, and behavioral responses may deviate from those predicted by standard rational-expectations models. This erosion

of trust can amplify market fragility, increase volatility, and weaken the transmission mechanisms of monetary and fiscal policy.

Financial markets are particularly sensitive to informational distortions. False narratives regarding corporate performance, sovereign stability, inflation trends, or geopolitical risks can trigger mispricing of assets, herding behavior, and abrupt shifts in capital flows. In extreme cases, coordinated disinformation campaigns may contribute to speculative bubbles, sudden market crashes, or liquidity shortages. Beyond short-term market disruptions, persistent exposure to disinformation can alter risk premia, reduce market depth, and discourage productive investment, thereby affecting long-run economic growth.

At the macroeconomic level, the impact of disinformation extends to key indicators such as inflation expectations, consumption and saving behavior, labor market participation, and investment dynamics. Misleading information about economic conditions or policy intentions may cause households and firms to postpone spending, accelerate precautionary saving, or reallocate resources inefficiently. Moreover, disinformation can weaken policy credibility, complicating the task of economic authorities and reducing the effectiveness of stabilization measures during periods of crisis.

Against this backdrop, the development of formal models that integrate disinformation into economic analysis has become increasingly important. Such models aim to capture how false or noisy information spreads, how it interacts with belief formation and trust, and how these processes translate into observable market and macroeconomic outcomes. By combining insights from economics, behavioral finance, network theory, and information economics, these frameworks provide tools for assessing systemic risks, quantifying welfare losses, and designing policy responses.

The objective of this study is to examine and synthesize existing and emerging models of disinformation's impact on trust, markets, and macroeconomic indicators. The analysis focuses on identifying

key transmission channels, evaluating model assumptions, and highlighting their implications for economic stability and policy design. Understanding these mechanisms is essential for improving risk monitoring, enhancing institutional resilience, and developing strategies to mitigate the economic consequences of disinformation in an increasingly interconnected and information-driven world.

Presentation of the main research material. Disinformation is conceptualized as deliberately false, misleading, or strategically distorted information that is disseminated with the intent to influence beliefs, expectations, or behavior of economic agents. Unlike random noise or unintentional misinformation, disinformation is characterized by its systematic nature, persistence, and potential coordination, which makes it particularly relevant for analyzing market dynamics and macroeconomic outcomes.

Trust is defined as the degree of confidence that economic agents place in information sources, institutions, markets, and policy authorities. In economic terms, trust reduces transaction costs, facilitates coordination, and supports the formation of stable expectations. It plays a central role in decision-making under uncertainty, affecting consumption, investment, saving behavior, and portfolio allocation. Within this framework, trust is treated as an endogenous variable that evolves over time in response to information quality, past experiences, and institutional credibility.

The interaction between disinformation and trust is modeled as a key transmission mechanism through which information shocks influence economic outcomes. Exposure to disinformation can weaken trust by increasing perceived uncertainty, distorting beliefs, and reducing the credibility of reliable information sources. As trust erodes, economic agents may rely more heavily on heuristics, peer behavior, or precautionary strategies, leading to deviations from rational expectations and amplifying market inefficiencies.

Furthermore, the framework distinguishes between different levels of trust, including interpersonal trust, institutional trust, and market

trust, each of which may respond differently to disinformation. These dimensions of trust influence economic behavior through distinct channels, such as risk perception, expectation formation, and responsiveness to policy signals. By explicitly incorporating these mechanisms, the conceptual framework establishes a foundation for formal modeling of how disinformation propagates through trust and ultimately affects financial markets and macroeconomic indicators.

Existing modeling frameworks are first evaluated for their suitability. Dynamic stochastic general equilibrium (DSGE) models are considered for analyzing macroeconomic transmission channels, as they allow disinformation to be introduced as information or expectation shocks that affect consumption, investment, inflation expectations, and policy credibility. In these models, trust can be incorporated as an endogenous state variable influencing agents' expectations and the effectiveness of monetary and fiscal policy.

Agent-based models (ABMs) are employed to capture heterogeneity, bounded rationality, and non-linear interactions among agents. These models are particularly useful for studying the diffusion of disinformation through social and financial networks, herding behavior, and abrupt regime shifts in markets. Trust in ABMs evolves dynamically based on agents' experiences and exposure to accurate or false information, allowing for the emergence of collective phenomena such as bubbles, crashes, or persistent mistrust.

Network and information diffusion models are used to represent the structure of communication channels and the propagation of disinformation across interconnected agents and institutions. These models help identify key nodes, amplification mechanisms, and tipping points where localized disinformation shocks generate systemic effects. They also allow for the analysis of coordinated disinformation campaigns and their impact on market sentiment.

Behavioral and learning-based models complement the analysis by relaxing the assumption of fully rational expectations. Adaptive learning, Bayesian updating with biased priors, and heuristic-based

decision rules are introduced to reflect how agents process noisy or deceptive information. Trust affects the weighting of information sources in belief updating, linking informational distortions directly to economic decisions.

Where existing frameworks are insufficient, a hybrid modeling approach is developed that integrates macroeconomic structure with agent-level interactions and information networks. This allows the simultaneous analysis of micro-level belief dynamics and macro-level outcomes. The resulting models provide a flexible and realistic representation of how disinformation affects trust, market behavior, and macroeconomic indicators, forming the basis for quantitative simulations and policy analysis (see Table 2.1, p. 36–37).

The classification presented in the table demonstrates that the economic impact of disinformation is inherently multidimensional and cannot be adequately captured by a single modeling framework. Different classes of models emphasize distinct mechanisms through which disinformation affects trust, markets, and macroeconomic indicators, highlighting the trade-offs between analytical tractability, realism, and policy relevance.

DSGE models provide a strong theoretical foundation and are well suited for analyzing aggregate dynamics and policy transmission mechanisms. Their ability to incorporate disinformation as expectation or credibility shocks makes them useful for studying macroeconomic stability and policy effectiveness. However, their reliance on representative or weakly heterogeneous agents limits their capacity to capture the diffusion of disinformation, social interactions, and abrupt regime changes driven by trust erosion.

Agent-based models offer a high degree of realism by explicitly modeling heterogeneous agents, bounded rationality, and non-linear interactions. They are particularly effective in capturing the endogenous evolution of trust and the emergence of phenomena such as herding, bubbles, and market crashes triggered by disinformation. The main limitation of ABMs lies in their

Table 2.1 – Models for Analyzing Disinformation's Economic Impact

Model Class	Core Features	Role of Trust	Disinformation Channel	Main Applications	Strengths	Limitations
1	2	3	4	5	6	7
DSGE Models (Dynamic Stochastic General Equilibrium)	Micro-founded, equilibrium-based, forward-looking agents	Endogenous or exogenous state variable affecting expectations and policy transmission	Modeled as information shocks, expectation distortions, or credibility shocks	Macroeconomic dynamics, monetary and fiscal policy analysis	Strong theoretical consistency, policy relevance	Limited heterogeneity, weak representation of network effects
Agent-Based Models (ABM)	Heterogeneous agents, bounded rationality, non-linear interactions	Evolves dynamically based on agent experiences and information quality	Explicit diffusion of disinformation through agent interactions	Market volatility, bubbles, crashes, trust erosion	Captures emergent behavior and non-linear dynamics	Calibration and validation can be complex
Network Models	Explicit network structure, node centrality, contagion effects	Trust embedded in link strength or node credibility	Spread of disinformation through social, media, or financial networks	Systemic risk, contagion, amplification mechanisms	Identifies key transmission nodes and tipping points	Often abstract, limited macro foundations

End of Table 2.1

1	2	3	4	5	6	7
Behavioral Models	Deviations from rational expectations, heuristics, biases	Influences belief updating and decision weights	Cognitive biases exploited by disinformation	Investor behavior, consumption and saving decisions	Realistic representation of human behavior	Less standardized, weaker aggregation to macro level
Learning Models (Adaptive / Bayesian)	Expectations updated over time using learning rules	Determines speed and accuracy of belief adjustment	Persistent bias in signals and priors	Expectation formation, inflation dynamics	Flexible and empirically testable	Sensitive to specification of learning rules
Information Economics Models	Asymmetric information, signaling, credibility	Trust linked to information reliability and reputation	Strategic manipulation of information	Market transparency, regulation, institutional credibility	Strong analytical clarity	Limited dynamic and network effects
Hybrid Models	Combination of macro structure, agents, and networks	Multi-level trust (institutional, market, interpersonal)	Joint diffusion and macro feedback loops	Integrated market – macro analysis	High realism and flexibility	Computationally intensive, complex interpretation

Source: compiled by the authors

computational complexity and challenges related to calibration, validation, and comparability across studies.

Network models play a critical role in identifying how disinformation propagates through interconnected systems and how local information shocks can generate systemic effects. By focusing on network topology, central nodes, and amplification mechanisms, these models are valuable for risk monitoring and early warning analysis. Nevertheless, their abstract nature often requires integration with macroeconomic or behavioral frameworks to generate policy-relevant conclusions.

Behavioral and learning-based models significantly enhance realism by relaxing the assumption of fully rational expectations. They capture how cognitive biases, heuristic decision-making, and adaptive belief updating interact with disinformation. While these models provide important insights into expectation formation and trust dynamics, they often face difficulties in aggregation and in linking micro-level behavior to macroeconomic outcomes in a consistent manner.

Information economics models contribute analytical clarity by emphasizing asymmetric information, signaling, and credibility. They are particularly useful for studying strategic disinformation and institutional trust. However, their static or partial-equilibrium nature limits their ability to address dynamic feedback effects and large-scale contagion processes.

Hybrid models represent the most promising direction for future research. By combining macroeconomic structure, agent heterogeneity, behavioral responses, and network-based diffusion, they allow for a more comprehensive analysis of how disinformation affects trust, markets, and macroeconomic indicators simultaneously. Despite their high computational and conceptual complexity, hybrid models offer the greatest potential for capturing real-world dynamics and informing effective policy interventions.

Overall, the evaluation suggests that model choice should be guided by the specific research question and policy objective. For

macroeconomic policy analysis, DSGE models remain essential, while agent-based and network models are indispensable for understanding trust dynamics and systemic risk. Integrating these approaches through hybrid frameworks appears to be the most effective strategy for advancing the economic analysis of disinformation.

Despite their analytical value, models used to study the economic effects of disinformation face a number of important limitations that constrain their applicability, interpretability, and policy relevance. These limitations arise from data availability, methodological assumptions, and the inherent complexity of information-driven phenomena.

A key limitation across all model classes is the measurement of disinformation and trust. Disinformation is difficult to observe directly, as it varies in intent, content, scale, and persistence. Proxy variables – such as media sentiment indices, social media activity, or survey-based trust measures – may only partially capture the true intensity and influence of disinformation. As a result, model outcomes can be sensitive to how these variables are defined and measured.

DSGE models are constrained by strong structural assumptions, including rational or near-rational expectations, equilibrium behavior, and representative agents. While trust can be incorporated as a state variable or shock, these models struggle to represent belief polarization, network-driven diffusion, and abrupt collapses in confidence. Consequently, DSGE-based results may underestimate the non-linear and systemic effects of disinformation, particularly during crises.

Agent-based models face challenges related to calibration, validation, and reproducibility. The high dimensionality of agent characteristics and behavioral rules often requires simplifying assumptions or ad hoc parameter choices. Empirical validation is difficult, as multiple parameter configurations may generate similar aggregate outcomes. This limits the comparability of results across studies and reduces their direct applicability for policy evaluation.

Network models are sensitive to assumptions about network structure, connectivity, and agent interactions. Real-world information networks are dynamic, multilayered, and often partially unobservable, whereas models typically rely on static or stylized representations. This can lead to biased estimates of contagion speed, amplification effects, and systemic importance of specific nodes.

Behavioral models introduce realism by incorporating cognitive biases and heuristics, but they often lack standardized microfoundations and face aggregation problems. Translating individual-level behavioral distortions into consistent macroeconomic outcomes remains challenging, limiting their integration into broader policy frameworks.

Learning-based models depend heavily on the specification of learning rules, information sets, and prior beliefs. Small changes in these assumptions can lead to significantly different dynamics, making results sensitive and sometimes unstable. Moreover, learning processes may not fully capture strategic manipulation of information characteristic of coordinated disinformation campaigns.

Information economics models typically focus on partial-equilibrium settings and strategic interactions under asymmetric information. While analytically elegant, they often abstract from dynamic feedback loops, large-scale diffusion, and macroeconomic adjustment processes, limiting their usefulness for studying economy-wide effects.

Finally, hybrid models, although offering the most comprehensive representation, are computationally intensive and difficult to interpret. Their complexity can obscure causal mechanisms, complicate communication with policymakers, and reduce transparency. Additionally, the integration of multiple modeling paradigms increases the risk of internal inconsistencies and overfitting.

In summary, while existing models provide valuable insights into the economic consequences of disinformation, their application requires careful consideration of underlying assumptions, data

limitations, and the specific context of analysis. Combining models, validating results across frameworks, and improving measurement of trust and information quality are essential steps toward more robust and policy-relevant conclusions.

Future research on the economic impact of disinformation should aim to address existing methodological and empirical limitations while reflecting the rapidly evolving information environment. Several promising directions can be identified for advancing both theoretical modeling and applied analysis.

First, there is a need for improved measurement of disinformation and trust. Future work should integrate high-frequency data from digital platforms, news analytics, and social media with traditional economic indicators and survey-based trust measures. Advances in natural language processing and machine learning can support the construction of more precise and timely indicators of disinformation intensity, narrative persistence, and credibility erosion.

Second, dynamic and endogenous trust formation should be further developed within macroeconomic models. Rather than treating trust as an exogenous shock or fixed parameter, future models should allow trust to evolve in response to information quality, institutional performance, and policy outcomes. Embedding these mechanisms into DSGE and semi-structural macro models would enhance their realism and policy relevance.

Third, greater emphasis should be placed on multi-layer network modeling. Real-world information transmission occurs across overlapping networks, including social media, traditional news, financial markets, and institutional communication channels. Future models should capture interactions across these layers and analyze how disinformation propagates and amplifies across different domains.

Fourth, hybrid modeling frameworks represent a key avenue for progress. Integrating agent-based dynamics, network diffusion, behavioral biases, and macroeconomic structure can

provide a unified framework capable of linking micro-level belief formation to macro-level outcomes. Advances in computational methods and increased data availability make such integration increasingly feasible.

Fifth, future research should focus on policy-oriented simulations and stress testing. Models of disinformation can be used to evaluate the effectiveness of countermeasures, such as transparency policies, fact-checking interventions, communication strategies of central banks, and platform-level regulations. Scenario analysis can help assess systemic vulnerabilities and the resilience of trust under coordinated disinformation campaigns.

Sixth, the role of expectation heterogeneity and belief polarization warrants deeper investigation. Future models should explicitly account for fragmented belief systems, echo chambers, and asymmetric information exposure, which can lead to persistent macroeconomic distortions and weaken policy transmission.

Finally, increased attention should be paid to cross-country and institutional heterogeneity. The economic impact of disinformation varies across institutional settings, levels of media freedom, and degrees of digital penetration. Comparative and international modeling approaches can improve the generalizability of results and support the design of context-specific policy responses.

Overall, future developments should move toward more data-driven, dynamic, and integrative modeling approaches. Such advances will be essential for understanding the systemic economic risks posed by disinformation and for designing effective strategies to protect trust, market stability, and macroeconomic performance in an increasingly information-driven economy.

To rigorously study the economic effects of disinformation, researchers can employ a combination of quantitative, survey-based, and network-oriented metrics. These metrics allow measurement of disinformation intensity, trust dynamics, market responses, and macroeconomic outcomes.

1. Metrics for Disinformation Intensity.

Volume of misinformation content: Number of posts, articles, or messages flagged as false or misleading on social media, news outlets, or messaging platforms.

Sentiment-based indices: Measures of emotional tone or polarity in media and social networks, indicating potentially manipulative content.

Misinformation reach and engagement: Number of views, shares, likes, and comments associated with disinformation content.

Disinformation propagation rate: Speed at which false narratives spread across networks or platforms, often measured in time to reach a certain number of users.

Content novelty and repetition: Frequency of recurring false narratives, reflecting persistence and reinforcement effects.

2. Metrics for Trust.

Consumer and investor confidence indices: Surveys such as the Consumer Confidence Index (CCI) or Investor Sentiment Index reflecting perception of economic conditions.

Institutional trust measures: Survey-based measures of trust in government, central banks, media, and financial institutions.

Market-based trust proxies: Measures derived from credit spreads, bond yields, or implied volatility indices that reflect market confidence.

Interpersonal trust surveys: Measures of social trust and perceived reliability of information sources.

Trust dynamics indicators: Changes in trust over time, derived from repeated surveys or longitudinal data.

3. Metrics for Market Impact.

Volatility indices (e.g., VIX): Indicators of market uncertainty sensitive to sudden informational shocks.

Asset price deviations: Differences between observed prices and fundamental values, potentially reflecting disinformation-driven mispricing.

Trading volumes and liquidity measures: Indicators of market activity and stress, which can spike during disinformation events.

Herding and clustering behavior: Statistical measures (e.g., cross-sectional standard deviation of returns) showing correlated market behavior.

4. Metrics for Macroeconomic Outcomes.

Consumption and savings rates: Changes in household behavior in response to perceived economic risk.

Investment and capital allocation: Shifts in firm-level or sectoral investment patterns influenced by information shocks.

Inflation and inflation expectations: Surveys of expected price changes, as disinformation can distort anticipatory behavior.

GDP growth and sectoral output: Macro-level economic performance indicators affected indirectly by trust erosion.

Employment and labor market participation: Changes reflecting precautionary or reactive behavior in response to perceived economic instability.

5. Network and Diffusion Metrics.

Centrality measures: Degree, betweenness, and eigenvector centrality to identify key nodes in disinformation propagation networks.

Clustering and modularity: Identifying echo chambers or communities where disinformation is amplified.

Propagation speed and reach: Temporal metrics capturing how quickly and widely disinformation spreads.

Influence scores: Quantifying the impact of specific actors, sources, or narratives on the network.

6. Combined / Derived Metrics.

Trust-adjusted market volatility: Market volatility weighted by survey-based or inferred trust levels.

Disinformation elasticity of expectations: Sensitivity of consumer, investor, or business expectations to changes in disinformation exposure.

Macro-financial risk indices: Composite indicators combining information diffusion, trust erosion, and market instability.

These metrics can be used individually or in combination to calibrate models, validate simulations, and monitor systemic risks.

They also facilitate the quantitative assessment of policy interventions, such as fact-checking initiatives, regulatory measures, or institutional communications, by measuring changes in trust, market behavior, or macroeconomic indicators.

Conclusions. The analysis of models, metrics, and research directions highlights several key insights regarding the economic impact of disinformation:

1. Disinformation is a systemic economic risk. Disinformation affects not only individual behavior but also collective expectations, market dynamics, and macroeconomic outcomes. Its effects propagate through trust networks, influencing consumption, investment, and financial market stability. The erosion of trust can amplify uncertainty, increase volatility, and reduce the effectiveness of economic policy.

2. No single model captures all dimensions. Different modeling frameworks – DSGE, agent-based, network, behavioral, learning-based, and hybrid models – emphasize different mechanisms and trade-offs. DSGE models provide rigorous macroeconomic foundations, while agent-based and network models excel at representing heterogeneity, diffusion, and non-linear effects. Hybrid approaches are the most promising for integrating micro-level belief dynamics with macroeconomic outcomes.

3. Trust is central. Trust functions as a key transmission mechanism through which disinformation affects economic decisions and market behavior. Both empirical and model-based analyses show that trust is dynamic, multi-dimensional, and highly sensitive to information quality and institutional credibility. Monitoring and modeling trust is therefore essential for understanding the broader economic impact of disinformation.

4. Metrics are diverse and complementary. A combination of quantitative, survey-based, market-based, and network metrics is necessary to measure disinformation intensity, trust levels, market responses, and macroeconomic effects. Composite and derived metrics, such as trust-adjusted volatility or disinformation elasticity

of expectations, can provide actionable insights for both researchers and policymakers.

5. Policy relevance and simulation potential. Modeling disinformation can inform policy interventions, including fact-checking, transparency measures, regulatory frameworks, and central bank communication strategies. Scenario-based simulations allow for the evaluation of systemic vulnerabilities and the effectiveness of measures designed to mitigate the economic consequences of disinformation.

6. Future research directions are clear. Key areas for future development include improved measurement of disinformation and trust, dynamic modeling of trust formation, multi-layer network integration, hybrid frameworks combining macro, micro, and behavioral dynamics, and comparative analyses across institutional and country contexts. Addressing these areas will enhance the robustness, realism, and policy relevance of economic models of disinformation.

7. Limitations must be acknowledged. All models face challenges related to data availability, parameter calibration, structural assumptions, and computational complexity. These limitations necessitate cautious interpretation of results and encourage the use of complementary approaches, cross-validation, and scenario analysis.

8. Overall, understanding and modeling the economic impact of disinformation requires an interdisciplinary, integrative approach that combines rigorous theory, empirical validation, and practical policy relevance. Disinformation is not merely a social or political issue – it is a measurable economic phenomenon with tangible effects on trust, markets, and macroeconomic stability.

References

1. Abdullahi, S. I. (2019). Measuring volatility linkage, clustering and sensitivity to external shocks in Nigerian stock index. *International Journal of Financial Services Management*, 9(4), 345–368.

2. Arendt, F., Haim, M., & Beck, J. (2019). Fake News, Warnhinweise und perzipierter Wahrheitsgehalt: Zur unterschiedlichen Anfälligkeit

für Falschmeldungen in Abhängigkeit von der politischen Orientierung [Fake news, warning messages, and perceived truth value: Investigating the differential susceptibility hypothesis related to political orientation]. Publizistik, 64, 181–204. https://doi.org/10/gfw6gt

3. Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic populism: Facebook pages of alternative news media and the corona crisis: A computational content analysis. ArXiv. https://arxiv.org/abs/2004.02566

4. Brody, D. C. (2022). Noise, fake news, and tenacious Bayesians. Frontiers in Psychology, 13, 797904.

5. Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1), 1–28. https://doi.org/10.18637/jss.v080.i01

6. Chung, M., & Jones-Jang, S. M. (2021). Red media, blue media, Trump briefings, and COVID-19: Examining how information sources predict risk preventive behaviors via threat and efficacy. Health Communication. Advance online publication. https://doi.org/10.1080/10410236.2021.1914386

7. Clarke, J., Chen, H., Du, D., & Hu, Y. J. (2020). Fake news, investor attention, and market reaction. Information Systems Research, 32(1), 35–52.

8. Davison, W. P. (1983). The third-person effect in communication. Public Opinion Quarterly, 47(1), 1–15. https://doi.org/10.1086/268763

9. Dice, M. (2017). The true story of fake news: How mainstream media manipulates millions. Mark Dice.

10. Faragó, L., Kende, A., & Krekó, P. (2020). We only believe in news that we doctored ourselves: The connection between partisanship and political fake news. Social Psychology, 51(2), 77–90. https://doi.org/10.1027/1864-9335/a000391

11. Hong, Y., Qu, B., Yang, Z., & Jiang, Y. (2023). The contagion of fake news concern and extreme stock market risks during the COVID-19 period. Finance Research Letters, 58, 104258.

12. Jackob, N., Quiring, O., & Schemer, C. (2017). Wölfe im Schafspelz? Warum manche Menschen denken, dass man Journalisten nicht vertrauen darf – und was das mit Verschwörungstheorien zu tun hat [Wolves in sheep's clothing? Why some people think you can't trust journalists – and the link to conspiracy theories]. In K. N. Renner, T. Schulz, & J. Wilke (Eds.),

Journalismus zwischen Autonomie und Nutzwert (pp. 225–250). Herbert von Halem Verlag.

13. Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. American Journal of Political Science, 60(4), 824–844. https://doi.org/10.1111/ajps.12234

14. Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. American Journal of Political Science, 60(4), 824–844. https://doi.org/10.1111/ajps.12234

15. Okon, P. E., Musa, J. H. T., & Oyesomi, K. (2021). Fake news circulation and regulation in Anglophone West Africa. *Indiana Journal of Humanities and Social Sciences*, 2(7), 35–49.

16. Pummerer, L., Böhm, R., Lilleholt, L., Winter, K., Zettler, I., & Sassenberg, K. (2022). Conspiracy theories and their societal effects during the COVID-19 pandemic. Social Psychological and Personality Science, 13(1), 49–59. https://doi.org/10.1177/19485506211000217

17. Talabi, F. O., Ugbor, I. P., Talabi, M. J., Ugwuoke, J. C., Oloyede, D., Aiyesimoju, A. B., & IkechukwuIlomuanya, A. B. (2022). Effect of a social media-based counselling intervention in countering fake news on COVID-19 vaccine in Nigeria. Health Promotion International, 37(2), daab140.

18. Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. William and Flora Hewlett Foundation. https://doi.org/10.2139/ssrn.3144139

2.2. INDICATORS OF INFORMATION RESILIENCE: OPERATIONALIZATION AND VALIDATION OF A COMPOSITE TRUST INDEX

Introduction. The contemporary information environment is undergoing profound transformation driven by digitalization, platformization, and the accelerating circulation of information across global networks. While these processes have significantly expanded access to information and enabled new forms of participation, they have also intensified structural vulnerabilities within information ecosystems. The rapid spread of misinformation and disinformation, declining confidence in traditional media, algorithmic amplification of polarizing content, and growing skepticism toward public institutions have collectively contributed to what is often described as a crisis of trust. In this context, the concept of information resilience has gained increasing relevance as a framework for understanding how societies cope with, adapt to, and recover from information-related shocks and disruptions.

Information resilience can be broadly defined as the capacity of individuals, communities, and institutional systems to maintain informed decision-making and social cohesion under conditions of information overload, uncertainty, and manipulation. Unlike purely technical approaches to information security, information resilience emphasizes cognitive, social, and institutional dimensions, highlighting the role of trust, media literacy, and adaptive governance. Trust, in particular, occupies a central position within this framework. It shapes how information is evaluated, which sources are considered credible, and how individuals respond to competing narratives. A deficit of trust can undermine democratic processes, weaken public compliance with policy measures, and exacerbate social fragmentation, while excessive or misplaced trust can increase susceptibility to manipulation and false information.

Despite widespread recognition of trust as a foundational element of information resilience, its empirical measurement remains

methodologically challenging. Trust is a multidimensional and context-dependent construct that operates across multiple levels, including interpersonal trust, trust in media organizations, trust in digital platforms, trust in experts, and trust in public institutions. Existing studies often rely on isolated indicators or single survey items, which limits their ability to capture the complexity of trust relationships within modern information environments. Moreover, the lack of standardized and validated measurement instruments hampers comparative analysis across countries, social groups, and time periods, reducing the practical applicability of research findings for policy development and resilience assessment.

This study responds to these challenges by proposing and empirically validating a composite Trust Index as an operational indicator of information resilience. The core premise of the paper is that trust, when conceptualized and measured as a structured combination of interrelated dimensions, can serve as a reliable proxy for assessing the robustness of information environments. Drawing on interdisciplinary theoretical foundations from sociology, communication studies, political science, and information science, the research conceptualizes trust not as a singular attitude but as a system of evaluative orientations toward information sources, intermediaries, and verification mechanisms.

The development of the composite Trust Index involves several critical steps. First, the study provides a theoretical justification for the selection of trust dimensions relevant to information resilience, including trust in traditional media, online media, public authorities, scientific and expert communities, and digital platforms. Second, these dimensions are operationalized through measurable indicators derived from survey data and existing empirical instruments. Third, the indicators are aggregated into a composite index using transparent methodological procedures, with careful consideration of normalization, weighting, and aggregation techniques. This approach seeks to balance methodological rigor with practical

usability, ensuring that the index can be applied across different empirical contexts.

A key contribution of this research lies in the validation of the proposed composite index. Validation is conducted through a combination of statistical techniques, including internal consistency testing, factor analysis, and construct validity assessment. In addition, the robustness of the index is examined across different socio-demographic groups and informational contexts to evaluate its sensitivity and generalizability. By systematically testing the index's empirical properties, the study aims to demonstrate that the composite Trust Index is not only theoretically grounded but also methodologically sound.

By operationalizing trust as a measurable and composite phenomenon, this paper advances the empirical study of information resilience and provides a practical tool for monitoring trust dynamics within information ecosystems. The findings have implications for researchers seeking to model information resilience, as well as for policymakers and practitioners involved in media regulation, strategic communication, and resilience-building initiatives. Ultimately, the study contributes to a more nuanced understanding of how trust functions as both a resource and a vulnerability in contemporary information societies, and how it can be systematically measured to support evidence-based decision-making. In addition to its analytical value, the measurement of trust plays an increasingly strategic role in the design and evaluation of public policy interventions aimed at strengthening information resilience. Governments, international organizations, and civil society actors are investing in counter-disinformation strategies, media literacy programs, and regulatory frameworks for digital platforms. However, the effectiveness of such interventions depends on the availability of reliable indicators that capture changes in trust over time and across different segments of the population. Without validated composite measures, policy responses risk being reactive, fragmented, or poorly targeted,

addressing symptoms rather than the underlying structural drivers of information vulnerability.

The growing emphasis on resilience-oriented governance further underscores the need for robust measurement tools. Resilience frameworks, originally developed in fields such as ecology and disaster risk management, have been increasingly applied to social and information systems. Within these frameworks, indicators serve not only as diagnostic instruments but also as mechanisms for learning and adaptation. A composite Trust Index can therefore function as an early-warning signal, revealing patterns of declining confidence or asymmetric trust distributions that may precede information crises. At the same time, such an index can be used ex post to assess recovery and adaptation following major information shocks, including election-related disinformation campaigns, public health crises, or geopolitical conflicts.

From a methodological perspective, the construction of composite indicators raises important questions concerning validity, reliability, and interpretability. Composite indices are inherently normative, as they involve decisions about which dimensions to include, how to operationalize them, and how to weight their relative importance. Critics often argue that aggregation may obscure meaningful variation or mask contradictory trends across components. This study directly engages with these critiques by explicitly documenting each step of the index construction process and by testing alternative model specifications. Sensitivity analyses are employed to examine how changes in weighting schemes or indicator selection affect the overall index, thereby enhancing transparency and analytical robustness.

Another challenge addressed in this research concerns the dynamic and context-specific nature of trust. Trust in information sources is shaped by cultural norms, historical experiences, media systems, and political structures. Consequently, indicators that perform well in one national or regional context may not be directly transferable to another. To mitigate this limitation, the proposed Trust Index is designed to be modular and adaptable, allowing researchers to adjust specific

components while maintaining a common conceptual core. This flexibility supports comparative research while preserving sensitivity to local information environments.

Furthermore, the study acknowledges that trust is not inherently positive or uniformly beneficial. High levels of trust in unreliable sources or opaque platforms may increase exposure to misinformation, while moderate skepticism can enhance critical information processing. Therefore, the Trust Index is not interpreted as a simple linear measure of information quality or democratic health. Instead, it is positioned as a diagnostic tool that captures patterns of trust distribution and alignment between users and credible information providers. This nuanced interpretation aligns with contemporary research emphasizing "calibrated trust" as a key element of information resilience.

By situating the composite Trust Index within broader debates on information governance, digital sovereignty, and societal resilience, the study seeks to bridge theoretical abstraction and empirical application. The index is intended not only as a measurement instrument but also as a conceptual lens through which trust-related dynamics in information ecosystems can be systematically analyzed. In doing so, the paper contributes to ongoing efforts to move beyond ad hoc metrics and toward standardized, validated indicators capable of informing long-term resilience strategies.

Presentation of the main research material. The conceptual research model developed in this study is grounded in the premise that information resilience can be analytically captured through structured patterns of trust within an information ecosystem. Rather than treating trust as a single latent attitude, the model conceptualizes it as a multidimensional construct embedded in interactions between individuals, information sources, intermediaries, and institutional frameworks. This approach allows for a systematic representation of how trust-related factors collectively shape the capacity of societies to withstand and adapt to information disturbances.

At the core of the conceptual model lies the assumption that information resilience emerges from the alignment between trust orientations and information system characteristics. The model distinguishes between several key domains of trust: trust in traditional media, trust in digital and social media platforms, trust in public institutions, trust in expert and scientific communities, and trust in information verification mechanisms. Each domain represents a distinct yet interrelated dimension that influences how information is accessed, evaluated, and acted upon. Together, these dimensions form the structural foundation of the proposed composite Trust Index.

The model further incorporates individual-level and contextual moderating factors that condition trust formation and its effects on information resilience. At the individual level, socio-demographic characteristics, media consumption patterns, and levels of media and digital literacy are assumed to influence trust distribution across information sources. At the contextual level, features of the information environment – such as media system structure, platform governance practices, regulatory regimes, and the prevalence of disinformation – are expected to shape aggregate trust dynamics. These factors do not function as direct components of the Trust Index but provide explanatory context for interpreting index variation.

Within the conceptual framework, trust operates as a mediating variable between information exposure and resilience outcomes. High levels of appropriately calibrated trust are expected to enhance information resilience by facilitating reliance on credible sources, supporting informed decision-making, and reducing vulnerability to manipulative content. Conversely, misaligned or polarized trust – characterized by high confidence in unreliable sources or systematic distrust of credible institutions – is hypothesized to weaken resilience and amplify informational risks. This mediating role positions trust as both a resource and a potential vulnerability within the information ecosystem.

The conceptual model also integrates outcome dimensions associated with information resilience, including resistance

to misinformation, adaptive information-seeking behavior, and sustained confidence in reliable information sources during periods of uncertainty or crisis. While these outcomes are not directly included in the composite Trust Index, they serve as external validation criteria for assessing its explanatory power. The relationship between the Trust Index and resilience outcomes is expected to be probabilistic rather than deterministic, reflecting the complex and non-linear nature of information processes.

Finally, the model is designed to support empirical operationalization and validation. Each trust dimension is linked to observable indicators derived from survey instruments and empirical datasets, enabling quantitative measurement and statistical testing. The conceptual clarity of the model ensures coherence between theoretical assumptions, measurement choices, and analytical procedures. By explicitly mapping the relationships between trust dimensions, contextual factors, and resilience outcomes, the conceptual research model provides a structured foundation for the development, validation, and application of the composite Trust Index as an indicator of information resilience.

The primary criterion guiding indicator selection is conceptual alignment with the defined dimensions of trust. Each empirical indicator is explicitly linked to one of the trust domains identified in the conceptual model, namely trust in traditional media, trust in digital and social media platforms, trust in public institutions, trust in expert and scientific communities, and trust in information verification mechanisms. This domain-based structure prevents conceptual overlap and ensures that each indicator captures a distinct aspect of trust relevant to information resilience. Indicators that could not be unambiguously assigned to a specific trust dimension were excluded in order to preserve analytical clarity.

A second criterion concerns empirical observability and measurement reliability. Preference is given to indicators that have been widely used and validated in previous large-scale surveys and

empirical studies, such as standardized trust questions and Likert-scale assessments. The use of established measurement instruments enhances comparability across studies and reduces the risk of measurement error. Where possible, indicators are selected to reflect stable evaluative orientations rather than short-term attitudes, thereby increasing their suitability for resilience-oriented analysis.

The study further emphasizes cross-contextual applicability as a key consideration in indicator selection. Information resilience is inherently shaped by diverse social, cultural, and political environments, which necessitates the use of indicators that can be meaningfully interpreted across different contexts. Accordingly, indicators are formulated in general terms and avoid references to specific media outlets, platforms, or institutions that may not exist or function similarly across settings. This approach supports the adaptability of the composite Trust Index for comparative and longitudinal research.

In addition, attention is paid to the balance between subjective and functional indicators of trust. While trust is primarily an attitudinal construct, its relevance to information resilience lies in its behavioral implications. Selected indicators therefore include both evaluative measures (e.g., perceived credibility or reliability of information sources) and functional measures (e.g., reliance on trusted sources for verification). This combination allows the index to capture not only stated confidence but also the practical role of trust in information-processing behavior.

The final set of indicators is subjected to preliminary statistical screening to assess internal consistency, variance, and redundancy. Indicators exhibiting low variance, high multicollinearity, or weak theoretical justification are excluded to enhance the parsimony and interpretability of the index. This step ensures that each retained indicator contributes meaningful information to the composite measure rather than inflating its complexity without analytical benefit.

Through this systematic selection and justification process, the study establishes a transparent and theoretically grounded

indicator framework for measuring trust as a core component of information resilience. The resulting set of empirical indicators provides a solid foundation for index construction, aggregation, and validation, thereby strengthening the overall explanatory capacity of the composite Trust Index.

Based on the conceptual model and the criteria outlined in the previous section, the following specific empirical indicators are proposed for operationalizing the composite Trust Index. Each factor is selected to reflect a distinct dimension of trust relevant to information resilience.

1. Trust in Traditional Media.

This dimension captures individuals' confidence in conventional news sources such as television, newspapers, and radio. Indicators include:

– Perceived credibility of major national news outlets (survey question: "How much do you trust the information provided by major national news channels / newspapers?"; Likert scale 1–5).

– Justification: Traditional media remain central sources of verified information, and their perceived credibility directly affects the resilience of the public against misinformation.

– Reliance on traditional media for news (survey question: "How often do you use TV, radio, or newspapers as your primary source of news?").

– Justification: Behavioral reliance complements attitudinal measures, capturing the practical role of trust in information-seeking behavior.

2. Trust in Digital and Social Media Platforms.

This dimension addresses confidence in online information intermediaries, including social networks and news aggregators. Indicators include:

– Perceived reliability of social media news feeds (survey question: "To what extent do you consider news on social media platforms to be reliable?").

– Justification: Social media are major conduits of both accurate and misleading information; understanding perceived reliability is key to assessing resilience.

– Frequency of cross-checking online information (survey question: "How often do you verify information found on social media with other sources?").

Justification: Reflects the functional aspect of trust, indicating adaptive behaviors that enhance resilience.

3. Trust in Public Institutions.

This dimension reflects confidence in government agencies, regulatory bodies, and other formal authorities. Indicators include:

– Trust in government-provided information (survey question: "How much do you trust official communications from government institutions?").

– Justification: Trust in institutions affects compliance with public guidance and influences susceptibility to misinformation.

– Perceived transparency and accountability of institutions (survey question: "How transparent do you consider public institutions in sharing accurate information?").

– Justification: Institutional transparency moderates trust and enhances resilience by enabling informed judgment.

4. Trust in Expert and Scientific Communities.

This dimension captures confidence in experts, researchers, and fact-checkers. Indicators include:

– Perceived credibility of scientific experts and researchers (survey question: "How much do you trust scientists and technical experts on topics of public concern?").

– Justification: Scientific trust is crucial during crises such as public health emergencies or climate-related events.

– Reliance on expert advice for decision-making (survey question: "How often do you consult expert sources before making decisions about current events?").

– Justification: Measures practical application of trust in expert knowledge.

5. Trust in Information Verification Mechanisms.

This dimension assesses confidence in tools and procedures used to assess information accuracy. Indicators include:

– Use of fact-checking websites and tools (survey question: "How often do you use online fact-checking services to verify information?").

– Justification: Active engagement with verification mechanisms is a core component of information resilience.

Confidence in the effectiveness of verification tools (survey question: "How confident are you that fact-checking tools provide accurate assessments?").

Justification: Trust in verification mechanisms reinforces adaptive information behavior and mitigates exposure to misinformation.

Table 2.2 – Summary of Empirical Indicator Structure

Trust Dimension	Indicator	Measurement Type
Traditional Media	Perceived credibility of major news outlets	Attitudinal (Likert 1–5)
Traditional Media	Reliance on traditional media	Behavioral frequency
Digital / Social Media	Perceived reliability of social media	Attitudinal (Likert 1–5)
Digital / Social Media	Frequency of cross-checking online information	Behavioral frequency
Public Institutions	Trust in government communications	Attitudinal (Likert 1–5)
Public Institutions	Perceived transparency/ accountability	Attitudinal (Likert 1–5)
Experts / Scientific Community	Perceived credibility of experts	Attitudinal (Likert 1–5)
Experts / Scientific Community	Reliance on expert advice	Behavioral frequency
Verification Mechanisms	Use of fact-checking tools	Behavioral frequency
Verification Mechanisms	Confidence in fact-checking tools	Attitudinal (Likert 1–5)

Source: compiled by the authors

The construction of the Composite Trust Index (CTI) represents a crucial step in operationalizing the conceptual model and transforming the selected empirical indicators into a unified measure of information resilience. The CTI is designed to integrate multiple dimensions of trust – traditional media, digital // /social media platforms, public institutions, experts / /scientific communities, and information verification mechanisms – into a single, interpretable metric that reflects overall confidence in the information ecosystem. The following methodological steps guide the construction of the index.

1. Indicator Normalization.

The first step in constructing the CTI involves normalizing the individual indicators to ensure comparability across dimensions with different measurement scales. Normalization transforms all indicators onto a common scale, typically ranging from 0 (lowest trust) to 1 (highest trust). In this study, the min-max normalization method is applied:

$$X_{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}},$$

where X_i is the raw score for a given respondent or observation, and X_{min} and X_{max} are the minimum and maximum observed values for that indicator. This method preserves relative differences while allowing aggregation across heterogeneous measures, including both attitudinal (Likert-scale) and behavioral indicators (frequency measures).

2. Weighting of Indicators.

Once normalized, indicators are weighted to reflect their relative importance within the composite index. In principle, several approaches can be applied:

– Equal weighting: Each indicator contributes equally to its dimension and to the overall CTI. This approach is simple, transparent, and commonly used when theoretical guidance on relative importance is limited.

– Empirical weighting: Factor analysis or principal component analysis (PCA) can be applied to determine weights based

on observed variance and the contribution of each indicator to the underlying latent construct.

– Expert-based weighting: Weights can be assigned based on expert assessment of the relative importance of different trust dimensions for information resilience.

For this study, a hybrid approach is proposed: equal weighting is applied within each trust dimension to balance the contribution of multiple indicators, while empirical weights derived from PCA are used across dimensions to capture their relative explanatory power in shaping overall trust patterns.

3. Aggregation of Indicators.

After normalization and weighting, indicators are aggregated at two levels:

1. Dimensional aggregation: Within each trust domain (e.g., traditional media), normalized indicators are combined using a weighted arithmetic mean:

$$T_d = \sum_{i=1}^{n_d} w_i X_{i,norm},$$

where T_d is the trust score for dimension d, w_i is the weight of indicator i within the dimension, and n_d is the number of indicators in that dimension.

2. Overall CTI aggregation: The dimension-level scores are then aggregated into a single composite index using weighted arithmetic or geometric mean:

$$CTI = \sum_{d=1}^{D} W_d T_d,$$

where W_d represents the weight of each trust dimension based on its contribution to information resilience, and D is the total number of trust dimensions (five in this study). The resulting CTI ranges from 0 (lowest overall trust) to 1 (highest overall trust), allowing for cross-sectional and longitudinal comparisons.

4. Validation and Robustness Checks.

The reliability and validity of the CTI are assessed using several complementary techniques:

- Internal consistency: Cronbach's alpha and composite reliability are calculated for each dimension and for the overall index to ensure that the indicators measure coherent constructs.
- Construct validity: Factor analysis is employed to confirm that indicators load appropriately onto their theoretical dimensions.
- Sensitivity analysis: Alternative weighting schemes (equal, empirical, expert-based) are tested to examine the robustness of the CTI.
- Convergent validity: The CTI is compared with external measures of information resilience outcomes, such as exposure to misinformation or adaptive information-seeking behavior, to evaluate its predictive validity.

5. Interpretation of the Composite Trust Index.

The resulting CTI provides a holistic measure of trust across multiple dimensions and serves as a proxy for the resilience of the information environment. Higher values indicate stronger confidence in credible information sources, institutions, and verification mechanisms, reflecting greater capacity to withstand and adapt to information shocks. Conversely, lower values indicate fragmented or misplaced trust, signaling potential vulnerabilities in the information ecosystem. By decomposing the CTI into dimension-specific scores, researchers can also identify areas of strength and weakness, informing targeted interventions for improving societal information resilience.

While the Composite Trust Index (CTI) provides a systematic and multidimensional measure of trust as a core component of information resilience, several limitations must be acknowledged to ensure a nuanced interpretation of the results.

One primary limitation is the complexity and abstraction of trust as a construct. Trust is inherently multidimensional, context-dependent,

and dynamic over time. While the CTI captures major dimensions – traditional media, digital platforms, public institutions, experts, and verification mechanisms – it cannot fully account for all situational, cultural, or cognitive factors that influence trust formation. For example, factors such as political polarization, historical experiences with institutions, or social network influences may shape trust in ways not fully reflected by the selected indicators.

The CTI relies on survey-based attitudinal and behavioral indicators, which are subject to well-known measurement biases. Respondent self-reporting may be affected by social desirability, recall error, or misunderstanding of questions. Moreover, some behavioral indicators (e.g., frequency of cross-checking information) may not fully capture the quality or effectiveness of trust-related actions. The limited availability of universally comparable indicators also constrains cross-cultural or cross-national applications.

The construction of the CTI involves subjective decisions regarding normalization, weighting, and aggregation. Although empirical and hybrid weighting methods enhance objectivity, these decisions can still influence the final index values and interpretation. Different weighting schemes could lead to divergent conclusions about relative trust levels across dimensions or populations. Similarly, the choice of aggregation method (arithmetic vs. geometric mean) affects sensitivity to extreme values or imbalances between dimensions.

Trust is dynamic and responsive to temporal events, such as crises, scandals, or policy interventions. The CTI, especially if based on cross-sectional data, provides a snapshot rather than a continuous measure of information resilience. Changes in public opinion or information environments may not be immediately reflected, limiting the index's capacity to capture short-term fluctuations or emerging vulnerabilities.

While the CTI offers a holistic measure, it cannot fully disentangle positive versus negative trust. High overall trust may not always indicate desirable resilience if it is directed toward unreliable sources,

and low trust may sometimes reflect critical skepticism rather than vulnerability. Consequently, interpretation requires careful attention to dimension-specific scores and external contextual information.

Finally, the applicability of the CTI depends on the availability and quality of survey or empirical data. Differences in sampling methods, question wording, and cultural interpretations of trust can affect comparability across populations or time periods. Generalizing findings to different socio-political or digital contexts should be done cautiously, and further validation may be required for each new setting.

In summary, while the Composite Trust Index provides a valuable tool for assessing multidimensional trust and its role in information resilience, its application should be complemented by careful contextual analysis, sensitivity testing, and triangulation with other data sources. Awareness of these limitations ensures responsible use and interpretation of the index in both research and policy contexts.

The Composite Trust Index (CTI) offers significant practical value for both researchers and policymakers seeking to strengthen information resilience within contemporary societies. By providing a multidimensional and empirically grounded measure of trust across media, institutions, experts, and verification mechanisms, the CTI enables the systematic identification of vulnerabilities in information ecosystems. Policymakers can use the index to monitor trust levels over time, detecting early signs of declining confidence that may signal susceptibility to misinformation or disinformation campaigns. Such monitoring supports the development of targeted interventions, including public communication strategies, media literacy programs, and transparency initiatives aimed at reinforcing trust in credible sources.

In addition, the CTI facilitates comparative analysis across different social groups, regions, or countries, allowing decision-makers to tailor policies to specific demographic or cultural contexts. For example, variations in trust in digital platforms versus traditional media may indicate the need for differentiated approaches to information verification or public engagement. By decomposing the index into its constituent

dimensions, practitioners can identify which aspects of the information environment are most robust and which require reinforcement, thereby optimizing resource allocation and strategic planning.

Beyond policy design, the CTI also serves as a diagnostic tool for media organizations, civil society actors, and digital platform providers. Media outlets can evaluate how their credibility is perceived relative to other information sources, informing editorial decisions and audience engagement strategies. Civil society organizations can assess the effectiveness of educational campaigns or fact-checking initiatives, using changes in trust scores as indicators of program impact. Digital platforms may employ the CTI to guide the development of user-oriented features that promote reliable information and support adaptive verification behaviors.

Furthermore, the index has applications in research and evaluation. By linking CTI scores with behavioral outcomes, such as information-seeking patterns or susceptibility to misinformation, researchers can test theoretical models of trust formation and information resilience. Longitudinal application of the index enables the study of trends over time, such as shifts in trust during crises or following major information interventions. In this way, the CTI not only measures the current state of trust but also supports evidence-based strategies for enhancing societal resilience to information disruptions.

Overall, the Composite Trust Index functions as both a measurement and management tool, bridging the gap between theoretical understanding of trust and practical efforts to strengthen the reliability, transparency, and resilience of information ecosystems. Its application can inform policy, guide media and platform strategies, and support ongoing research, contributing to a more informed, adaptive, and resilient society.

The practical utility of the CTI extends beyond immediate policy and media applications, offering a strategic framework for long-term societal planning in the face of evolving information challenges. In environments characterized by rapid technological

change and increasing information complexity, the index can serve as an early-warning system, signaling emerging vulnerabilities in public trust before they translate into widespread misinformation or social fragmentation. By tracking shifts in trust patterns across time, institutions can proactively adjust communication strategies, develop resilience-oriented education initiatives, and strengthen the credibility of key information intermediaries.

In addition, the CTI provides a valuable tool for evaluating the effectiveness of interventions aimed at improving information resilience. For instance, governments implementing transparency campaigns or media literacy programs can use the index to assess whether these efforts lead to measurable improvements in public confidence across specific trust dimensions. Similarly, media organizations and digital platforms can employ the CTI to test the impact of initiatives such as fact-checking services, algorithmic adjustments, or credibility labeling on user trust and engagement. By linking index scores to behavioral and attitudinal outcomes, stakeholders gain actionable insights into which strategies successfully enhance resilience and which require refinement.

The index also facilitates comparative and cross-cultural research, enabling the examination of how trust operates within different socio-political and technological contexts. Understanding these variations is crucial for designing policies that are contextually appropriate and culturally sensitive, particularly in societies where historical experiences or institutional frameworks shape trust differently. Moreover, the CTI can inform collaborative efforts among governments, international organizations, and civil society actors to strengthen global information resilience, providing a standardized metric for monitoring progress, sharing best practices, and coordinating multi-level interventions.

Finally, the CTI contributes to fostering a more adaptive and informed society by promoting awareness of trust dynamics and the importance of credible information sources. By highlighting areas

of weakness in public confidence, it encourages both institutions and individuals to engage in critical evaluation of information, to seek verification, and to prioritize reliability over volume or speed of content consumption. Over time, widespread application of the CTI can support the development of more resilient information ecosystems, enhancing the capacity of societies to navigate uncertainty, resist manipulation, and make informed decisions in complex informational environments.

Beyond immediate monitoring and evaluation, the Composite Trust Index offers the potential to guide strategic decision-making in complex information environments. In times of crisis – such as public health emergencies, natural disasters, or geopolitical conflicts – the index can help authorities and media organizations identify which segments of the population are more vulnerable to misinformation and which sources are most trusted. This enables targeted communication strategies that maximize the effectiveness of public messaging, ensuring that critical information reaches audiences through channels they consider credible. In addition, the CTI can support scenario planning by simulating how shifts in trust across different dimensions may influence public responses to policy measures or social initiatives.

For digital platform governance, the CTI provides actionable insights into user trust dynamics. Social media companies, news aggregators, and other online intermediaries can analyze CTI patterns to design features that promote reliable information, improve transparency, and encourage critical engagement with content. For example, the index may reveal low confidence in platform verification tools, prompting the development of more effective fact-checking integrations or user education campaigns. Similarly, the CTI can inform algorithmic decisions aimed at balancing engagement with credibility, helping platforms mitigate the spread of misinformation while maintaining user trust.

The index also has implications for education and public awareness programs. By highlighting areas where trust in expert communities

or verification mechanisms is low, policymakers and educators can design initiatives to strengthen media literacy, scientific reasoning, and critical thinking. This, in turn, contributes to long-term societal resilience by fostering a population that is better equipped to evaluate information, question unreliable sources, and actively engage in informed decision-making.

Furthermore, the CTI can serve as a tool for longitudinal research and policy evaluation, providing a standardized measure that tracks changes in trust over time and across contexts. This allows governments, international organizations, and academic researchers to assess the impact of interventions, regulatory reforms, or societal events on information resilience. Comparative application across countries or regions can reveal best practices and structural factors that contribute to high levels of trust, offering insights for replication in diverse informational and cultural environments.

In sum, the Composite Trust Index functions not merely as a measurement tool but as a strategic instrument for enhancing societal resilience, guiding policy interventions, and informing responsible information governance. Its multidimensional approach allows stakeholders to understand the complex interplay between trust, behavior, and information exposure, providing a foundation for evidence-based decision-making in both short-term crisis response and long-term resilience-building efforts.

Conclusions. This study has developed a comprehensive framework for assessing information resilience through the operationalization of a Composite Trust Index (CTI). Trust, as a multidimensional and context-sensitive construct, is central to understanding the capacity of individuals, communities, and institutions to navigate complex information environments. By integrating trust in traditional media, digital and social media platforms, public institutions, expert and scientific communities, and information verification mechanisms, the CTI provides a holistic measure of the robustness and adaptability of information ecosystems.

The conceptual model underpinning this research emphasizes the interplay between trust dimensions, individual behaviors, and contextual factors. It recognizes that trust is both a resource and a potential vulnerability: appropriately calibrated trust enables reliance on credible information and informed decision-making, whereas misaligned trust increases susceptibility to misinformation and social fragmentation. The model also accounts for moderating influences, such as socio-demographic characteristics, media consumption patterns, digital literacy, and structural features of information environments. This theoretical grounding ensures that the CTI captures not only attitudinal dimensions but also behavioral and functional aspects of trust, providing a nuanced and actionable representation of information resilience.

The selection of empirical indicators was guided by conceptual relevance, empirical observability, and cross-contextual applicability. Indicators were chosen to reflect both evaluative and behavioral facets of trust, ensuring that the index captures the practical role of trust in information processing and verification. Through systematic normalization, weighting, and aggregation, the individual indicators were combined into a coherent composite index. The resulting CTI enables cross-sectional and longitudinal comparisons, while its dimensional decomposition allows for the identification of specific strengths and vulnerabilities within the information ecosystem. Validation procedures, including internal consistency testing, factor analysis, and sensitivity analyses, support the reliability and interpretability of the index, providing confidence in its utility for research and policy applications.

Despite its advantages, the CTI has several limitations that must be carefully considered. Trust is inherently dynamic, context-dependent, and influenced by cultural, political, and technological factors, which the index cannot fully capture. Survey-based indicators are susceptible to measurement biases, and aggregation and weighting decisions introduce normative elements that may influence outcomes.

Furthermore, high levels of overall trust are not inherently positive if directed toward unreliable sources, highlighting the need for careful interpretation of both aggregate and dimension-specific scores. These limitations underscore the importance of contextual analysis, triangulation with other data sources, and ongoing refinement of the index to enhance its applicability across diverse settings.

From a practical perspective, the CTI offers significant value for policymakers, media organizations, digital platform providers, and researchers. It functions as an early-warning tool, enabling the identification of declining or misaligned trust before it manifests in widespread misinformation or behavioral vulnerabilities. The index can guide targeted interventions, support media literacy initiatives, inform platform governance and algorithmic design, and provide a basis for evaluating the effectiveness of resilience-building programs. Comparative and longitudinal applications of the CTI also allow for evidence-based monitoring of trends in trust, facilitating cross-cultural research and the identification of best practices. Ultimately, the index contributes to a deeper understanding of the mechanisms underpinning information resilience and offers a practical framework for enhancing societal capacity to navigate complex and uncertain information environments.

In conclusion, the development and operationalization of the Composite Trust Index represent a meaningful advance in the empirical study of information resilience. By bridging conceptual clarity, methodological rigor, and practical applicability, the CTI provides a robust tool for assessing trust dynamics, guiding policy interventions, and fostering resilient information ecosystems. Future research should focus on refining the index, expanding its cross-cultural applicability, and integrating it with broader measures of social and cognitive resilience. Such efforts will further strengthen our ability to understand and enhance the adaptive capacity of societies in the face of increasingly complex informational challenges. The development and operationalization of the Composite Trust

Index (CTI) provide not only a methodological tool for measuring trust but also a strategic framework for enhancing information resilience across societal, institutional, and technological domains. By integrating multiple dimensions of trust – traditional media, digital platforms, public institutions, expert and scientific communities, and verification mechanisms – the CTI captures the complex interplay of attitudes, behaviors, and contextual factors that underpin the capacity to navigate and adapt to information challenges. Its multidimensional design allows stakeholders to identify both strengths and vulnerabilities within information ecosystems, offering a granular understanding that can inform targeted interventions.

From a policy perspective, the CTI offers several actionable insights. Governments and regulatory bodies can utilize the index to monitor the health of public trust in real time, identifying early signs of declining confidence that may exacerbate vulnerability to misinformation or polarizing narratives. By analyzing dimension-specific scores, policymakers can tailor communication strategies to address specific gaps – for example, enhancing transparency and responsiveness in public institutions, supporting expert communication channels, or promoting the visibility of credible media sources. In addition, the index can inform the design and evaluation of media literacy programs, public awareness campaigns, and regulatory interventions, providing empirical evidence for the effectiveness of policies aimed at strengthening societal resilience.

For media organizations, the CTI highlights the importance of credibility, reliability, and audience engagement in shaping trust. News outlets can use dimension-specific insights to assess how their reporting is perceived relative to other information sources, enabling more effective editorial strategies and audience outreach. Emphasizing transparency in reporting, verification of facts, and clear communication of uncertainty can enhance trust scores and contribute to a more resilient public information environment. Media organizations can also collaborate with fact-checking initiatives

and digital platforms to amplify the visibility of reliable information and promote informed public discourse.

For digital platforms, the CTI provides critical guidance for designing user-centered interventions that support resilience. Low trust in verification tools or platform governance features may signal the need for improved fact-checking integrations, algorithmic transparency, or content moderation policies. By analyzing behavioral indicators such as cross-checking frequency and reliance on expert sources, platforms can identify opportunities to foster critical engagement and mitigate the spread of misinformation. The index also supports platform evaluation over time, allowing developers to measure the impact of algorithmic adjustments, feature updates, or community guidelines on user trust and resilience.

Beyond immediate applications, the CTI has long-term strategic implications. It enables longitudinal monitoring of trust dynamics, capturing trends and shifts resulting from crises, societal events, or technological changes. This capacity is critical in increasingly complex and interconnected information environments, where shifts in trust can have rapid and wide-ranging consequences for public opinion, democratic processes, and societal cohesion. By linking CTI scores to behavioral outcomes and resilience measures, stakeholders can develop predictive models, scenario analyses, and targeted interventions that anticipate vulnerabilities rather than reactively addressing their consequences.

Finally, the CTI contributes to a broader understanding of societal adaptation and learning in the information age. By highlighting areas of weakness and strength, the index encourages both institutions and individuals to engage in practices that enhance critical thinking, verify information, and prioritize credible sources. Over time, such engagement can foster a culture of calibrated trust, where individuals are neither overly credulous nor excessively skeptical, and institutions are more accountable, transparent, and responsive. In this sense, the CTI functions not only as a measurement tool but also as a catalyst

for systemic improvements in information governance, public awareness, and societal resilience.

In conclusion, the Composite Trust Index represents a significant advancement in the empirical study of information resilience, bridging theory, methodology, and practice. Its application provides actionable insights for policymakers, media, and digital platforms, offering a robust framework for monitoring, guiding, and enhancing trust in information ecosystems. Future research should focus on refining the index through cross-cultural validation, integrating dynamic and context-sensitive indicators, and exploring its predictive capacity for resilience outcomes. By doing so, the CTI can continue to serve as both a diagnostic and strategic instrument, enabling societies to navigate the challenges of the modern information landscape with greater awareness, adaptability, and resilience.

References

1. Holling, C. S., Walker, B. Resilience Defined. Entry prepared for the Internet Encyclopaedia of Ecological Economics. Int. Soc. Ecol. Econ. 2003.

2. Frankenberger, T., Constas, M., Nelson, S., Starr, L. Current Approaches to Resilience Programming among Nongovernmental Organisations; 2020 Conference Paper 7; International Food Policy Research Institute: Washington, DC, USA, 2014.

3. Oregon Seismic Safety Policy Advisory Commission (OSSPAC). The Oregon Resilience Plan: Reducing Risk and Improving Recovery for the Next Cascadia Earthquake and Tsunami; The Commission: Salem, OR, USA, 2013.

4. Bach, C., Birkmann, J., Kropp, J., Olonscheck, M., Setiadi, N., Vollmer, M., Walther, C. Assessing Vulnerability to Heat Waves and Heavy Rainfall at a Community Level. Pract. Exp. Civ. Prot. 2014, 11, 1–164.

5. Keating, A., Campbell, K., Mechler, R., Michel-Kerjan, E., Mochizuki, J., Kunreuther, H., Bayer, J., Hanger, S., McCallum, I., See, L., et al. Operationalizing Resilience against Natural Disaster Risk: Opportunities, Barriers, and a Way Forward; Zurich Flood Resilience

Alliance: Zurich, Switzerland, 2014; Available online: http: //opim.wharton.upenn.edu/risk/library/ZAlliance-Operationalizing-Reslience.pdf

6. Sorg, L., Medina, N., Feldmeyer, D., Sanchez, A., Vojinovic, Z., Birkmann, J., Marchese, A. Capturing the multifaceted phenomena of socioeconomic vulnerability. Nat. Hazards 2018, 1–26.

7. Cutter, S. L., Boruff, B. J., Shirley, W. L. (2003). Social Vulnerability to Environmental Hazards. Social Science Quarterly, 84(2): 242–261. https://doi.org/10.1111/1540-6237.8402002

8. Asadzadeh, A., Kötter, T., Salehi, P., Birkmann, J. (2017). Operationalizing a concept: The systematic review of composite indicator building for measuring community disaster resilience. International Journal of Disaster Risk Reduction, 25: 147–162. https://doi.org/10.1016/j.ijdrr.2017.09.015

9. Pfefferbaum, B. J., Reissman, D. B., Pfefferbaum, R. L., Klomp, R. W., Gurwitch, R. H. (2007). Building Resilience to Mass Trauma Events. Handbook of Injury and Violence Prevention, eds. Doll L. S, Bonzo S. E., Sleet D. A., Mercy J. A. (Springer US, Boston, MA), pp. 347–358. https://doi.org/10.1007/978-0-387-29457-5_19

10. Walpole, E. H., Loerzel, J., Dillard, M. (2021). A Review of Community Resilience Frameworks and Assessment Tools: An Annotated Bibliography (National Institute of Standards and Technology). https://doi.org/10.6028/NIST.TN.2172

11. Dillard, M. K. (2018). Developing an Assessment Methodology for Community Resilience. RESILIENCE: The 2nd International Workshop on Modeling of Physical, Economic, and Social Systems for Resilience Assessment. Vol. 2.

12. USDHS (2016). Mitigation Framework Leadership Group (MitFLG) Draft Concept Paper: Draft Interagency Concept for Community Resilience Indicators and National-Level Measures (U. S. Department of Homeland Security, Washington, D. C.).

13. Glinz, M. (2007). On Non-Functional Requirements. 15th IEEE International Requirements Engineering Conference (RE 2007) (IEEE, Delhi), pp. 21–26.

14. Sherrieb, K., Norris, F. H., Gale, S. (2010). Measuring Capacities for Community Resilience. Social Indicators Research, 99(2): 227–247. https://doi.org/10.1007/s11205-010-9576-9

15. Smith, R., Simard, C., Sharpe, A. (2001). A Proposed Approach to Environment and Sustainable Development Indicators Based on Capital. (The National Round Table on the Environment and the Economy's Environment and Sustainable Development Indicators Initiative).

16. Basu, S., Berkowitz, S. A., Phillips, R. L., Bitton, A., Landon, B. E., Phillips, R. S. (2019). Association of Primary Care Physician Supply With Population Mortality in the United States, 2005–2015. JAMA Internal Medicine, 179(4): 506. https://doi.org/10.1001/jamainternmed.2018.7624

CHAPTER 3.
AI TECHNOLOGIES AND PLATFORMS FOR COUNTERING DISINFORMATION

3.1. NATURAL LANGUAGE PROCESSING AND LARGE LANGUAGE MODELS: ARCHITECTURES, DATASETS, METRICS

Introduction. Natural Language Processing (NLP) represents a cornerstone of artificial intelligence, aiming to bridge the gap between human communication and computational understanding. It encompasses a wide array of techniques that allow machines to comprehend, interpret, generate, and respond to human language in a way that is both meaningful and contextually relevant. The evolution of NLP over the past decades has been remarkable, transitioning from rule-based systems to statistical models, and ultimately to modern deep learning approaches that leverage massive amounts of data and computational power. At the heart of this evolution lies the development of Large Language Models (LLMs), which have transformed the landscape of natural language understanding and generation, enabling unprecedented capabilities in tasks ranging from machine translation and summarization to conversational agents and creative content generation.

The advent of LLMs is closely intertwined with breakthroughs in deep learning architectures, particularly neural networks, which have demonstrated remarkable ability to capture complex patterns and dependencies in text. These models, typically trained on billions of parameters, are designed to learn rich representations of language, encoding semantic, syntactic, and contextual information in ways that were previously unattainable. Architectures such as Transformers,

which underpin most modern LLMs, leverage mechanisms like self-attention to model long-range dependencies efficiently, allowing models to process and generate coherent text over extended sequences. This architectural innovation has not only improved performance across a wide range of NLP benchmarks but has also opened avenues for transfer learning, where pre-trained models can be fine-tuned on specific tasks with limited labeled data, thus democratizing access to high-performance language technologies.

Central to the development and evaluation of both NLP systems and LLMs are the datasets on which these models are trained and tested. High-quality datasets serve as the foundation for learning linguistic patterns, capturing diverse language phenomena, and ensuring model generalization across various contexts and domains. These datasets range from curated corpora of formal text, such as news articles and scientific publications, to large-scale web crawls that capture informal and colloquial usage. The sheer scale and diversity of modern datasets are crucial for training LLMs, as they enable models to acquire broad linguistic knowledge and adapt to multifaceted language tasks. However, dataset selection and curation also raise significant challenges, including biases, ethical concerns, and data quality issues, all of which can directly influence model behavior and downstream applications.

Evaluation metrics constitute another fundamental aspect of NLP and LLM research, providing standardized methods to quantify model performance across different tasks. Traditional metrics, such as BLEU, ROUGE, and METEOR, have long been used to assess translation, summarization, and other generation tasks. More recently, specialized metrics, including perplexity, accuracy, F1-score, and human evaluation scores, have become standard in measuring both the fluency and factual correctness of model outputs. Selecting appropriate metrics is critical, as it informs model development, highlights limitations, and guides the alignment of model behavior with human expectations. Furthermore, the rise of LLMs has spurred the development of new evaluation paradigms, focusing not only on task-specific performance

but also on aspects like robustness, reasoning capabilities, fairness, and alignment with ethical standards.

The convergence of advanced architectures, expansive datasets, and rigorous evaluation frameworks has positioned NLP and LLMs at the forefront of contemporary AI research. They offer transformative potential across industries, including healthcare, finance, education, and entertainment, by automating language-intensive tasks, enhancing human-computer interaction, and enabling new forms of content creation. Despite these advancements, the field continues to face significant challenges, such as model interpretability, energy consumption, ethical considerations, and the mitigation of biases. Addressing these challenges requires a holistic understanding of model architectures, the characteristics of training data, and the limitations of evaluation metrics, ensuring that future developments in NLP and LLMs are not only technologically sophisticated but also socially responsible and aligned with human values.

In conclusion, Natural Language Processing and Large Language Models represent a dynamic intersection of linguistic theory, computational innovation, and data-driven learning. Their architectures, datasets, and metrics collectively define the capabilities and limitations of modern language technologies, shaping the way humans interact with machines and the ways in which machines understand human language. As research in this domain continues to accelerate, it is imperative to develop robust, ethical, and efficient NLP systems that harness the power of large-scale language models while addressing the complexities and nuances inherent in human communication. This field stands as a testament to the remarkable progress of artificial intelligence and its profound implications for society, technology, and human knowledge.

Presentation of the main research material. Natural Language Processing (NLP) has evolved through several distinct phases, each reflecting advances in computational methods, linguistic theory, and the availability of data. Early NLP systems, dating back

to the 1950s and 1960s, were primarily rule-based. These systems relied on manually encoded grammatical rules, lexicons, and heuristics to process text. While pioneering at the time, rule-based approaches were limited in scalability, brittle in handling linguistic variability, and often unable to generalize across domains or languages. The 1980s and 1990s saw the emergence of statistical methods, driven by the availability of larger corpora and advances in probabilistic modeling. Techniques such as Hidden Markov Models (HMMs) and n-gram language models enabled more flexible and data-driven approaches to tasks like part-of-speech tagging, speech recognition, and machine translation. These statistical models marked a significant shift from rigid rules toward empirically grounded learning, laying the foundation for modern NLP.

The advent of deep learning in the 2010s revolutionized NLP by introducing neural network-based architectures capable of capturing complex semantic and syntactic patterns. Early neural models, such as word embeddings (e.g., Word2Vec and GloVe), provided dense vector representations of words, capturing semantic relationships and enabling more sophisticated processing. However, these models were often limited by fixed-length context windows and their inability to capture long-range dependencies in text. The introduction of the Transformer architecture in 2017, through the landmark paper "Attention is All You Need", fundamentally changed this paradigm. Transformers utilize self-attention mechanisms to model relationships between all tokens in a sequence simultaneously, allowing for unprecedented scalability and context modeling. This architecture became the foundation for the development of Large Language Models (LLMs), which are typically pre-trained on vast corpora and fine-tuned for specific downstream tasks.

Large Language Models, such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), T5 (Text-to-Text Transfer Transformer), and LLaMA (Large Language Model Meta AI), exemplify the transformative

impact of these architectures. BERT introduced bidirectional context modeling, enabling models to consider both left and right context when learning word representations, which significantly improved performance on tasks like question answering and sentiment analysis. GPT models, developed by OpenAI, focus on autoregressive generation, producing coherent and contextually relevant text by predicting the next token in a sequence. T5 reframed NLP tasks in a unified text-to-text format, allowing a single model to perform translation, summarization, and classification tasks without task-specific architectures. Each of these models represents a step toward more generalizable and capable NLP systems, with billions of parameters and training on diverse datasets spanning multiple domains and languages.

Datasets play a critical role in the success of LLMs, as they provide the linguistic knowledge and diversity necessary for model learning. Curated datasets such as Wikipedia, Common Crawl, BookCorpus, and OpenWebText have been widely used for pre-training, capturing a broad spectrum of language use from formal to informal text. Additionally, domain-specific datasets, including biomedical texts, legal documents, and technical manuals, enable models to specialize in particular areas of expertise. The construction and selection of datasets must balance diversity, quality, and ethical considerations, as biases and inaccuracies in the training data can propagate into model outputs. Moreover, large-scale datasets require sophisticated preprocessing, deduplication, and filtering to ensure that models learn meaningful patterns rather than memorizing noise or harmful content.

Evaluation metrics are equally essential in assessing the capabilities of NLP systems and LLMs. For traditional NLP tasks, metrics such as BLEU, ROUGE, METEOR, accuracy, precision, recall, and F1-score provide quantitative measures of model performance. For generative models, newer metrics focus on text coherence, factual accuracy, and alignment with human preferences. Human evaluation remains a gold standard, particularly for creative or subjective tasks,

highlighting the limitations of automated metrics in fully capturing language quality. Beyond performance metrics, emerging research emphasizes evaluating models for robustness, fairness, interpretability, and energy efficiency, reflecting the growing awareness of the broader societal impacts of LLM deployment.

The convergence of sophisticated architectures, vast datasets, and comprehensive evaluation metrics has positioned NLP and LLMs as central drivers of innovation across multiple industries. From virtual assistants and automated customer support to scientific research and creative writing, these models demonstrate the capacity to transform human-computer interaction and augment human capabilities. However, the field continues to face challenges, including mitigating biases, ensuring interpretability, reducing environmental impact, and aligning models with ethical and societal norms. Addressing these challenges requires interdisciplinary collaboration between computer scientists, linguists, ethicists, and domain experts, ensuring that the development of NLP technologies is responsible, sustainable, and aligned with human values.

In summary, Natural Language Processing and Large Language Models represent a convergence of linguistic insight, computational innovation, and data-driven learning. The historical progression from rule-based systems to statistical models and finally to deep learning and Transformers underscores the rapid evolution of the field. Modern LLMs, powered by expansive datasets and evaluated with rigorous metrics, exemplify the potential of AI to understand, generate, and interact with human language at unprecedented scales. As the field continues to advance, it is imperative to combine technical excellence with ethical responsibility, ensuring that these transformative technologies contribute positively to society while navigating the inherent complexities of language and communication.

The rapid evolution of Large Language Models (LLMs) has been driven largely by architectural innovations that enable effective learning and generation of natural language. While many LLMs

share the foundational Transformer architecture, variations in design, training objectives, and scale have led to distinct capabilities. Among the most influential architectures are BERT, GPT, T5, and LLaMA, each exemplifying different approaches to representation learning and language generation.

1. BERT (Bidirectional Encoder Representations from Transformers).

BERT, introduced by Google in 2018, is a bidirectional encoder that reads text in both directions simultaneously. Unlike previous unidirectional models that only consider context from left-to-right or right-to-left, BERT leverages masked language modeling (MLM), predicting randomly masked tokens in a sequence based on both preceding and succeeding words. This bidirectional approach allows BERT to capture deep contextual relationships between words, enhancing performance on tasks requiring nuanced understanding, such as question answering, sentiment analysis, and named entity recognition.

Key Features:

- Bidirectional context modeling.
- Pre-training on large corpora (Wikipedia, BookCorpus).
- Fine-tuning for downstream tasks.
- Strong performance on NLP benchmarks like GLUE and SQuAD.

Limitations:

- Not optimized for text generation (primarily used for understanding).
- Requires separate fine-tuning for different tasks.

2. GPT (Generative Pre-trained Transformer).

Developed by OpenAI, GPT models represent autoregressive language models, which generate text by predicting the next token in a sequence. GPT is unidirectional in its context modeling, focusing on left-to-right generation. The model is pre-trained on massive datasets using next-token prediction and can be fine-tuned or used in a zero-shot/one-shot setting for various downstream tasks. GPT's architecture has scaled dramatically from GPT-1 (117M

parameters) to GPT-4 (trillions of parameters in underlying training models), demonstrating the impact of scale on performance.

Key Features:

- Autoregressive generation for coherent text output.
- Pre-training enables few-shot, zero-shot, and fine-tuned applications.
- Widely used in chatbots, content creation, coding assistance, and creative writing.

Limitations:

- Can generate plausible but factually incorrect outputs.
- Unidirectional context may limit some understanding tasks.

3. T5 (Text-to-Text Transfer Transformer).

T5, introduced by Google in 2020, represents a unified text-to-text approach, where all NLP tasks are reformulated as text generation tasks. For instance, a classification task becomes "Input: [text]. Output: [label]". This design enables the same model architecture to handle translation, summarization, question answering, and classification, simplifying multi-task learning and transfer learning. T5 is based on an encoder-decoder Transformer architecture, combining the strengths of bidirectional encoding with autoregressive decoding.

Key Features:

- Unified text-to-text framework.
- Encoder-decoder architecture for both understanding and generation.
- Flexible for multi-task training and fine-tuning.

Limitations:

- Computationally intensive for large-scale deployment.
- Requires careful prompt formulation for task performance.

4. LLaMA (Large Language Model Meta AI).

LLaMA, developed by Meta AI, is designed as a highly efficient and scalable LLM, emphasizing parameter efficiency while maintaining competitive performance. Unlike extremely large models such as GPT-4, LLaMA models are optimized to achieve strong results

on benchmarks with fewer parameters, making them more accessible for research and deployment. LLaMA employs the Transformer decoder architecture with optimizations for training speed, memory usage, and generalization.

Key Features:

- Parameter-efficient design.
- Strong performance on NLP benchmarks with smaller model sizes.
- Designed for research accessibility and reproducibility.

Limitations:

- Fewer publicly available models and datasets compared to GPT or BERT.
- Primarily focused on English and limited multilingual capability in smaller versions.

Table 3.1 – Comparative Overview

Model	Architecture	Context	Main Strength	Typical Use Cases
BERT	Encoder-only	Bidirectional	Understanding and classification	QA, Sentiment Analysis, NER
GPT	Decoder-only	Left-to-right	Text generation	Chatbots, Content Creation, Coding
T5	Encoder-decoder	Bidirectional + autoregressive	Unified multi-task learning	Summarization, Translation, Classification
LLaMA	Decoder-only	Left-to-right	Efficiency and scalability	Research, Benchmarking, General NLP

Source: compiled by the authors

LLaMA, which stands for Large Language Model Meta AI, is a family of advanced large language models developed by Meta AI. The LLaMA series represents Meta's effort to create high-performance, flexible, and research-friendly foundational models that can be studied,

fine-tuned, and deployed across a wide range of natural language processing tasks. Introduced in February 2023, LLaMA provided an open-source alternative to proprietary models like GPT, making state-of-the-art language modeling more accessible to researchers and developers. Early versions of LLaMA ranged from 7 billion to 65 billion parameters, achieving competitive performance even against much larger models such as GPT-3. Since then, the LLaMA family has expanded to include newer generations, including LLaMA 2, LLaMA 3, and LLaMA 4, with increasing scale, efficiency, and multimodal capabilities. These models are designed as foundation models, serving as pre-trained networks that can be adapted to numerous downstream tasks through fine-tuning.

The core architecture of LLaMA is a decoder-only Transformer, optimized for autoregressive language generation. The model leverages self-attention layers to capture relationships between tokens in a sequence, multi-head attention to model different types of dependencies simultaneously, and feed-forward layers to process and transform representations between attention blocks. Positional information is incorporated using Rotary Positional Embeddings (RoPE), which improve the handling of long sequences, while normalization layers like RMSNorm stabilize training and enhance generalization. Because LLaMA is a decoder-only model, it excels in generative tasks, including text completion, creative writing, and code generation, though it can also be fine-tuned for classification or other NLP tasks.

LLaMA comes in a wide range of model sizes, balancing performance, computational cost, and accessibility. Small models, such as those with 1–8 billion parameters, are suitable for experimentation and edge deployment. Mid-range models, like the 70-billion-parameter version, provide strong performance for research and commercial tasks, while the largest models, such as LLaMA 3.1 with 405 billion parameters, offer state-of-the-art capabilities in generation, reasoning, and long-context understanding. More recent LLaMA generations

incorporate techniques such as mixture-of-experts (MoE), activating only a subset of parameters during inference, which enables extremely large models to run efficiently. In addition, newer LLaMA variants include multimodal capabilities, allowing them to process text alongside images, enabling tasks such as image captioning, visual question answering, and multimodal reasoning.

The training of LLaMA models involves massive datasets containing trillions of tokens, drawn from diverse sources to ensure broad linguistic knowledge and cross-domain generalization. Larger versions include multilingual data, enabling cross-language understanding and generation. Training such large models requires advanced techniques like model parallelism and optimized attention mechanisms to manage memory and speed up computation.

One of LLaMA's major contributions is its research-friendly accessibility. While initial releases had some restrictions on usage, subsequent versions provide licensing that allows fine-tuning, experimentation, and controlled commercial deployment. This openness has fostered a growing ecosystem of derivative models, fine-tuning techniques like LoRA, quantization methods, and experimental deployments across domains ranging from scientific research to creative writing.

LLaMA has been applied in a wide array of tasks, including natural language generation, summarization, conversational AI, multilingual translation, scientific analysis, and vision-language tasks for multimodal models. Despite its advantages, LLaMA faces challenges, including questions about the transparency of training data, competition with other leading LLMs, and the need for careful evaluation of reasoning and factual accuracy.

In summary, LLaMA is a major family of large language models from Meta AI, featuring a decoder-only Transformer architecture, a range of model sizes from research-friendly to extremely large, multimodal and multilingual capabilities, and a research-focused open framework. Its flexibility, efficiency, and performance make

it a foundational model for both academic research and real-world applications, while ongoing development continues to push the boundaries of scale, efficiency, and multimodal capabilities in AI.

GPT, short for Generative Pre-trained Transformer, is a family of large language models developed by OpenAI that has become one of the most influential lines of models in natural language processing. GPT is designed as an autoregressive, decoder-only Transformer, optimized for text generation and capable of performing a wide variety of language tasks with little to no task-specific fine-tuning. Its development marked a significant milestone in NLP, demonstrating that scaling up model size and training data, combined with self-supervised pre-training, can produce models that generalize across a vast array of language tasks.

The first GPT model, introduced in 2018, was based on the Transformer architecture proposed by Vaswani et al. in 2017. GPT-1 demonstrated that pre-training on a large corpus of text followed by task-specific fine-tuning could outperform previous state-of-the-art methods on multiple NLP benchmarks. With 117 million parameters, GPT-1 was relatively modest in scale, but it validated the potential of unsupervised pre-training combined with supervised fine-tuning.

GPT-2, released in 2019, represented a dramatic leap in model size and capability, with up to 1.5 billion parameters. GPT-2 was trained on the WebText dataset, containing over 8 million web pages, and showcased unprecedented fluency in text generation, able to produce coherent, contextually relevant paragraphs across a wide range of prompts. Its ability to generate realistic and coherent text raised discussions about AI safety, misuse, and ethical deployment, as OpenAI initially restricted the full release due to concerns about disinformation and automated content generation.

GPT-3, released in 2020, scaled the model to 175 billion parameters, trained on hundreds of billions of tokens from a combination of Common Crawl, web pages, books, and other diverse sources. GPT-3 introduced the concept of few-shot and zero-shot

learning, allowing the model to perform new tasks simply by providing examples or natural language instructions, without explicit task-specific fine-tuning. This demonstrated that very large-scale pre-trained models could generalize in ways that smaller models could not.

GPT-4, the latest iteration (released in 2023–2024), continues this trajectory of scaling and capability improvement. While exact model sizes and architecture details are proprietary, GPT-4 incorporates multimodal inputs (text and images) and demonstrates advanced reasoning, problem-solving, and code generation capabilities. GPT-4's architecture builds on the decoder-only Transformer foundation, with improvements in training efficiency, context handling, alignment with human preferences, and reinforcement learning from human feedback (RLHF).

GPT models use a decoder-only Transformer architecture, optimized for autoregressive generation:

– Self-Attention Layers: Capture dependencies across the input sequence, enabling the model to consider both local and long-range relationships.

– Multi-Head Attention: Allows the model to attend to multiple aspects of the sequence in parallel, improving representation learning.

– Feed-Forward Networks: Apply nonlinear transformations to the attention outputs, enabling complex feature extraction.

– Layer Normalization and Residual Connections: Stabilize training, improve convergence, and allow very deep architectures.

– Positional Encoding: Provides the model with information about the position of tokens in a sequence, which is essential for maintaining word order and context.

As a decoder-only model, GPT predicts the next token in a sequence given the previous tokens, making it highly effective for text completion, story generation, question answering, and even coding tasks.

GPT models are trained on massive and diverse datasets to ensure broad linguistic coverage and generalization:

– WebText (GPT-2): Web pages filtered for high-quality content.

– Common Crawl: Extensive web crawls containing informal and formal text.

– BooksCorpus and Wikipedia: Structured text sources to improve factual knowledge and context understanding.

Other curated datasets: Large-scale proprietary data sources to enhance domain diversity.

The combination of scale, quality filtering, and diversity enables GPT models to generate coherent, context-aware text across domains, languages, and tasks.

GPT models demonstrate remarkable generalization, supporting tasks such as:

Text Generation: Producing coherent paragraphs, creative writing, summaries, and translations.

• Question Answering: Providing fact-based or context-aware responses to queries.

• Conversational AI: Powering chatbots, virtual assistants, and interactive dialogue systems.

• Code Generation: Writing and debugging code, especially in GPT-3.5 and GPT-4 with coding-focused fine-tuning.

• Instruction Following: Understanding and executing tasks described in natural language without explicit programming.

GPT models are used across industries, including customer service automation, education, content creation, scientific research, and software development.

A key innovation in GPT-3 and beyond is the ability to perform tasks with few-shot, one-shot, or zero-shot prompting:

• Zero-Shot: The model performs a task without any examples, relying solely on instructions.

• One-Shot: The model is given a single example before performing the task.

• Few-Shot: The model is provided with several examples, improving performance without task-specific fine-tuning.

This approach allows GPT models to generalize to tasks not explicitly seen during training, making them highly versatile.

Despite its capabilities, GPT has several limitations:

- Factual Inaccuracy: The model can generate plausible but incorrect or misleading information ("hallucinations").
- Bias and Ethical Concerns: GPT can reflect societal biases present in training data.
- Compute and Environmental Costs: Training and deploying such large models require significant computational resources and energy.
- Context Window Limitations: Earlier GPT models are limited by the number of tokens they can process at once; although GPT-4 supports extended context, it still has practical limits.

Addressing these challenges requires careful model alignment, fine-tuning, filtering, and ethical deployment practices.

GPT models have fundamentally transformed NLP and AI. From GPT-1's initial experiments in unsupervised pre-training to GPT-4's multimodal reasoning capabilities, the GPT family demonstrates that scaling model size, leveraging large datasets, and using autoregressive Transformers can produce highly versatile language models. GPT models excel in text generation, reasoning, and instruction-following tasks, and have become the foundation for commercial and research applications worldwide. While challenges like bias, factual accuracy, and energy consumption remain, GPT represents a milestone in AI, illustrating the transformative potential of large-scale, pre-trained language models.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a landmark development in the field of natural language processing introduced by Google in 2018. Unlike previous language models that processed text in a unidirectional manner, BERT was designed to understand context bidirectionally, meaning it considers both the left and right context of a word simultaneously. This innovation allowed BERT to capture deeper semantic

relationships in text and dramatically improved performance on a wide range of NLP tasks, including question answering, sentiment analysis, named entity recognition, and text classification.

The architecture of BERT is based on the encoder portion of the Transformer, which enables it to generate rich contextual embeddings for every token in a sequence. These embeddings capture nuanced relationships between words, phrases, and sentences, allowing the model to understand meaning beyond simple word co-occurrence. BERT is pre-trained on two primary tasks: masked language modeling (MLM) and next sentence prediction (NSP). In masked language modeling, a certain percentage of tokens in a sentence are randomly masked, and the model must predict these tokens based on the surrounding context. This forces the model to develop a deep understanding of how words relate to each other in both directions. In next sentence prediction, the model learns to determine whether a given sentence logically follows another, which improves its understanding of sentence-level relationships and coherence.

BERT is pre-trained on massive text corpora, primarily Wikipedia and BookCorpus, providing it with a broad knowledge of language, grammar, and general world knowledge. Once pre-trained, BERT can be fine-tuned for specific downstream tasks by adding task-specific layers, which allows it to adapt to specialized applications with relatively small labeled datasets. This fine-tuning process has been instrumental in making BERT highly versatile across different NLP domains while maintaining strong generalization capabilities.

One of BERT's key contributions is its ability to achieve state-of-the-art performance on several benchmark datasets upon its release, including the GLUE benchmark, SQuAD (Stanford Question Answering Dataset), and SWAG (Situations With Adversarial Generations). These achievements demonstrated the effectiveness of bidirectional context modeling and pre-training in comparison to earlier unidirectional models such as traditional LSTMs or word embedding-based methods.

Despite its groundbreaking capabilities, BERT is not optimized for generative tasks. Since it uses an encoder-only architecture, it is primarily focused on understanding and representation rather than text generation. Models designed for generation, such as GPT, complement BERT by providing autoregressive capabilities for coherent text output. Additionally, the standard BERT model has limitations in processing very long sequences due to fixed input size constraints, which has led to the development of variants like Longformer and BigBird that extend BERT's applicability to longer contexts.

BERT has inspired a wide range of derivative models, including RoBERTa, ALBERT, DistilBERT, and Electra, each building upon BERT's encoder framework with modifications to improve efficiency, scalability, or performance on specific tasks. These variants have become integral to the NLP ecosystem, demonstrating the foundational role BERT has played in shaping modern approaches to language understanding.

In summary, BERT represents a paradigm shift in natural language processing, introducing bidirectional context understanding through Transformer encoders and demonstrating the power of large-scale pre-training combined with task-specific fine-tuning. Its ability to generate rich contextual embeddings, achieve state-of-the-art results, and inspire numerous derivative models underscores its central importance in both research and practical NLP applications. BERT remains a cornerstone of modern language modeling, particularly for tasks that require deep understanding of text rather than generative output.

Beyond its original architecture and pre-training objectives, BERT's significance in natural language processing lies in its ability to serve as a universal feature extractor for text. By generating contextual embeddings for each token, BERT allows downstream models to leverage rich semantic representations without requiring task-specific feature engineering, which was a major bottleneck in earlier NLP pipelines. These embeddings encode not just the identity of individual words but also their nuanced meaning depending

on surrounding context, capturing polysemy (words with multiple meanings) and syntactic relationships in a way that static embeddings like Word2Vec or GloVe could not. This bidirectional understanding enables more accurate reasoning in tasks such as entity recognition, coreference resolution, and reading comprehension, where the meaning of a word often depends on distant parts of a sentence or paragraph.

BERT's pre-training process emphasizes large-scale self-supervised learning, making it efficient in leveraging unannotated text from massive corpora. Masked language modeling encourages the model to predict hidden words based on context from both directions, promoting deeper language comprehension, while next sentence prediction allows BERT to capture relationships between sentences, an ability crucial for natural language inference and question-answering systems. The combination of these objectives gives BERT an edge in tasks requiring both fine-grained token-level understanding and broader sentence-level reasoning.

Since its release, BERT has inspired numerous architectural and methodological variants aimed at addressing limitations in size, speed, and adaptability. RoBERTa (Robustly Optimized BERT Pretraining Approach) demonstrated that removing the next sentence prediction task and increasing training data and batch size could improve performance across NLP benchmarks. ALBERT (A Lite BERT) reduced the number of parameters through factorized embedding parameterization and cross-layer parameter sharing, making the model more memory-efficient without sacrificing accuracy. DistilBERT introduced a smaller, faster version of BERT using knowledge distillation, enabling deployment in resource-constrained environments while retaining most of BERT's capabilities. These derivatives illustrate the flexibility and foundational role of BERT in advancing NLP research and applications.

In practical applications, BERT has become a backbone for a wide range of tasks in both industry and research. It is commonly used in search engines for query understanding and document ranking, in chatbots and virtual assistants to interpret user intent, and

in information extraction systems for automatically identifying entities, relations, and key facts from text. In addition, BERT-based models have been applied to specialized domains, such as biomedical text mining and legal document analysis, where understanding nuanced terminology and context is critical.

Despite its impact, BERT does have limitations. Its encoder-only architecture makes it less suited for generative tasks, such as creative text generation or dialogue synthesis, where decoder or encoder-decoder models like GPT or T5 are more appropriate. Furthermore, BERT's input sequence length is fixed, typically to 512 tokens, which can limit its ability to process very long documents without truncation or hierarchical processing. Training and fine-tuning BERT also require substantial computational resources, which can be a barrier for smaller research teams or commercial deployments.

Overall, BERT represents a transformative advancement in NLP, shifting the focus from task-specific feature engineering to large-scale pre-training and bidirectional contextual understanding. Its design principles, pre-training strategies, and versatile embeddings have not only set new performance benchmarks but also established a foundation for a broad ecosystem of derivative models. BERT continues to be a central tool in NLP research and industry applications, providing both a methodological framework and practical capabilities for understanding natural language at scale.

T5, which stands for Text-to-Text Transfer Transformer, is a highly influential large language model developed by Google Research and introduced in 2020. Unlike models that separate tasks into distinct architectures or formats, T5 proposes a unified text-to-text framework, where all natural language processing tasks – ranging from translation and summarization to question answering and classification – are reformulated as text generation problems. In this approach, both the input and output are treated as text sequences, allowing a single model architecture to handle a wide variety of tasks without requiring task-specific modifications. This innovation simplifies model design,

facilitates multi-task learning, and leverages the power of pre-training to generalize across diverse NLP problems.

T5 is built on a full Transformer architecture with both encoder and decoder components. The encoder reads the input sequence and produces contextualized embeddings for each token, while the decoder generates output tokens autoregressively, attending to both the encoded input and the tokens it has generated so far. This encoder-decoder design allows T5 to excel at tasks that require both understanding of input context and coherent generation of output text, bridging the gap between comprehension-focused models like BERT and generation-focused models like GPT.

The pre-training of T5 involves a massive dataset called Colossal Clean Crawled Corpus (C4), a cleaned and filtered version of web crawl data containing hundreds of billions of tokens. During pre-training, T5 is optimized using a span-corruption objective, in which contiguous spans of text are masked and the model is tasked with generating the missing content. This strategy is more sophisticated than simple masked token prediction, as it forces the model to understand broader context and generate coherent spans, improving its ability to perform generation tasks effectively.

T5's text-to-text paradigm is highly versatile. Tasks such as machine translation are represented by providing input like "translate English to French: [sentence]" and generating the translated sentence as output. Summarization is framed as "summarize: [document]" with the corresponding summary as output, and classification can also be expressed in text form, for example "classify sentiment: [sentence]" with the output being "positive" or "negative." This unified format enables multi-task learning and makes it easier to apply transfer learning across tasks, since the model does not require separate heads or architectures for each problem.

Since its introduction, T5 has demonstrated state-of-the-art performance on a wide range of benchmarks, including GLUE, SuperGLUE, SQuAD, CNN/Daily Mail summarization, and

multilingual translation tasks. Its success has inspired numerous variants, such as mT5, a multilingual version capable of processing over 100 languages, and Flan-T5, which incorporates instruction tuning to improve zero-shot and few-shot generalization on unseen tasks. These adaptations highlight T5's flexibility and scalability, making it applicable in both research and industrial settings.

T5 excels in tasks requiring both comprehension and generation, making it particularly effective for summarization, question answering, translation, and any scenario where output text must be generated coherently from an input context. However, as a large encoder-decoder model, T5 is computationally intensive and requires significant resources for training and deployment, which can pose challenges for resource-constrained environments. Despite this, its unified approach simplifies model maintenance and deployment compared to systems that require multiple task-specific models.

In summary, T5 represents a paradigm shift in NLP modeling by unifying diverse tasks under a single text-to-text framework. Its encoder-decoder Transformer architecture, combined with large-scale pre-training on C4 and the span-corruption objective, enables it to achieve high performance across both understanding and generation tasks. T5's versatility, adaptability to instruction tuning, and ability to generalize across tasks make it a foundational model in modern NLP, bridging the strengths of models focused on comprehension and those optimized for generative tasks. It continues to serve as a powerful tool for researchers and practitioners seeking a flexible and general-purpose language model capable of handling a wide spectrum of natural language challenges.

While BERT, GPT, T5, and LLaMA have each advanced natural language processing in significant ways, they also exhibit inherent limitations shaped by their architectures, training data, and design objectives. Understanding these constraints is critical for responsible deployment, model selection, and ongoing research in large language models.

BERT, as an encoder-only model, excels at understanding and representing text but is not designed for text generation. Its bidirectional context modeling provides strong comprehension capabilities, yet it cannot produce coherent sequences of text without additional architecture or decoding mechanisms. BERT also has limitations in handling very long input sequences, as its typical maximum token length of 512 constrains processing of long documents or multi-paragraph contexts. Additionally, training and fine-tuning BERT require substantial computational resources, which can be prohibitive for smaller organizations. Like all large models, BERT inherits biases present in its pre-training datasets, potentially reflecting demographic, cultural, or gendered biases in downstream applications.

GPT, in contrast, is optimized for generative tasks and demonstrates impressive fluency in text generation. However, it has notable weaknesses in factual accuracy, frequently producing plausible but incorrect or misleading information, a phenomenon known as "hallucination." GPT's autoregressive, unidirectional design can also limit nuanced understanding of context compared to bidirectional models like BERT. Furthermore, GPT models are highly resource-intensive, both in training and inference, making deployment expensive. GPT also amplifies biases from its training data, and without careful alignment, it may produce outputs that are offensive, harmful, or ethically problematic. Context windows, although extended in GPT-4, remain limited, which can affect tasks requiring reasoning over very long documents.

T5, with its encoder-decoder architecture and text-to-text paradigm, is highly versatile and capable of both understanding and generation. Nevertheless, it is computationally demanding, as training and inference involve processing both input and output sequences through large Transformer networks. While the unified text-to-text format simplifies task adaptation, it also introduces challenges in task specification: the model's performance is sensitive to prompt formulation, task

instructions, and fine-tuning methods. T5 models may still hallucinate information in generative tasks and are affected by biases in the pre-training corpus. In addition, despite being able to handle multiple tasks, extremely long inputs or outputs can strain memory and processing capacity, limiting scalability for some applications.

LLaMA, designed to provide high-performance language modeling with greater parameter efficiency, also exhibits limitations. As a decoder-only Transformer, it is optimized for autoregressive generation and may struggle with certain comprehension-focused tasks unless fine-tuned. LLaMA inherits biases and errors from its large-scale pre-training data, similar to other LLMs, and it may produce inaccurate or harmful outputs without proper alignment. While smaller and more efficient than models like GPT-3 or GPT-4, larger LLaMA models still require significant computational resources for training and inference. Early versions also had limitations in multilingual coverage and transparency regarding training data, and even the multimodal extensions are constrained by the same architectural and resource limitations inherent in massive Transformer models.

In summary, each of these prominent language models – BERT, GPT, T5, and LLaMA – offers distinct strengths aligned with their architectures, but all face trade-offs in terms of computational cost, bias, factual accuracy, context length, and task suitability. BERT excels in comprehension but cannot generate text; GPT excels in generative fluency but may hallucinate and amplify biases; T5 is versatile but computationally heavy and sensitive to task formulation; and LLaMA provides efficiency and scalability but shares challenges in alignment, bias, and resource requirements. Recognizing these limitations is essential for responsible research, deployment, and continued innovation in natural language processing and large language models (see Table 3.2, p. 99).

The performance and capabilities of Large Language Models (LLMs) are heavily influenced by the datasets they are trained on and the metrics used to evaluate them. These two aspects are foundational,

shaping not only the quality of model outputs but also their ethical implications, robustness, and generalizability across domains.

1. Datasets.

Datasets in NLP serve two primary purposes: pre-training and fine-tuning/evaluation. Pre-training datasets provide the large-scale raw data necessary for models to learn general linguistic patterns, while fine-tuning datasets focus on task-specific learning, enabling models to adapt to concrete applications.

1.1. Pre-training DatasetsPre-training datasets are typically massive often containing billions of words across multiple domains. These datasets allow LLMs to develop a broad understanding of language and world knowledge. Key examples include:

• Wikipedia: Provides structured and high-quality encyclopedic text, widely used for general pre-training.

• Common Crawl: Web-crawled data representing diverse language use, from news to forums, blogs, and social media.

Table 3.2 – Summarizing the limitations of BERT, GPT, T5, and LLaMA

Model	Main Limitations
BERT	Encoder-only architecture; cannot generate text; limited input length (usually 512 tokens); computationally expensive for training and fine-tuning; susceptible to biases in pre-training data.
GPT	Autoregressive, unidirectional context; prone to hallucinations and factual errors; high computational and memory requirements; biases and ethical concerns; limited context window despite extensions in GPT-4.
T5	Encoder-decoder architecture is computationally heavy; sensitive to prompt and task formulation; can hallucinate in generation tasks; struggles with very long inputs or outputs; affected by pre-training biases.
LLaMA	Decoder-only model optimized for generation, may struggle on comprehension tasks; inherits biases from training data; larger models require significant computational resources; limited transparency in some versions; multilingual and multimodal capabilities are constrained.

Source: compiled by the authors

• BookCorpus: Collections of books used to provide narrative and literary context, improving models' understanding of long-range dependencies and storytelling.

• OpenWebText: An open-source approximation of OpenAI's WebText, including web content filtered for quality.

Challenges in pre-training datasets include biases, noise, and data quality, as large-scale web data may contain misinformation, toxic language, or overrepresented demographic or cultural content. Dataset curation, deduplication, and filtering are critical steps to mitigate these risks and ensure model generalization.

1.2. Fine-tuning and Benchmark Datasets. After pre-training, LLMs are fine-tuned on smaller, task-specific datasets to optimize performance for particular applications:

• GLUE (General Language Understanding Evaluation): A collection of classification tasks, including sentiment analysis, linguistic acceptability, and semantic similarity.

• SQuAD (Stanford Question Answering Dataset): A reading comprehension dataset for extractive question answering.

• CoNLL-2003: Focused on named entity recognition, used to evaluate entity extraction performance.

• CNN/Daily Mail: A dataset for abstractive summarization of news articles.

• SuperGLUE: An extension of GLUE with more challenging language understanding tasks.

Fine-tuning datasets are smaller but high-quality, enabling models to specialize and achieve state-of-the-art performance on benchmarks that reflect practical NLP applications.

2. Evaluation Metrics.

Evaluation metrics provide quantitative and qualitative measures of model performance. Metrics are chosen based on the task type – classification, generation, or understanding – and are critical for model development, comparison, and benchmarking.

2.1. Classification Metrics.

- Accuracy: Percentage of correctly predicted labels.
- Precision, Recall, and F1-Score: Measure correctness, completeness, and harmonic mean of predictions, especially useful for imbalanced datasets.

2.2. Generation Metrics.

- BLEU (Bilingual Evaluation Understudy): Measures n-gram overlap for translation or text generation tasks.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Measures overlap of n-grams and sequences for summarization tasks.
- METEOR: Considers synonymy and stemming, providing more nuanced evaluation than BLEU.
- Perplexity: Evaluates how well a language model predicts a sequence, with lower perplexity indicating better predictive power.

2.3. Human-Centered and Task-Specific Evaluation.

- Human Evaluation: Involves assessing fluency, coherence, factual accuracy, and alignment with human preferences, especially critical for creative generation and dialogue.
- Task-Specific Scores: For example, Exact Match (EM) in SQuAD evaluates the percentage of perfectly answered questions, while Rouge-L captures longest common subsequence overlap in summarization.
- Robustness, Fairness, and Bias Metrics: Evaluate model behavior under adversarial inputs, demographic fairness, and reduction of harmful outputs.

2.4. Emerging Evaluation Paradigms. With the rise of LLMs, there is increasing interest in holistic evaluation frameworks that go beyond task-specific metrics. These include:

- Alignment Metrics: Evaluate whether model outputs conform to ethical guidelines, factual correctness, and human intent.
- Multimodal Evaluation: For models integrating text, images, or audio, specialized metrics assess cross-modal reasoning and generation.

• Efficiency Metrics: Measure computational cost, inference speed, and energy consumption, reflecting practical deployment considerations.

The future of large language models is poised to be shaped by continued architectural innovation, scale optimization, multimodal integration, efficiency improvements, and ethical alignment. While models like BERT, GPT, T5, and LLaMA have demonstrated remarkable capabilities in understanding, generating, and reasoning with language, their limitations highlight opportunities for next-generation models to evolve in both functionality and societal impact.

One prominent trend is the integration of multimodal capabilities, allowing models to process not only text but also images, audio, and video. LLaMA and GPT-4 have already begun to explore this direction, but future models will likely extend multimodal reasoning to more complex tasks, such as understanding scientific diagrams, video sequences, or multi-language content in real-time. This could enable more sophisticated human-computer interaction, creative content generation, and knowledge discovery across domains.

Another key trajectory is scaling model efficiency and accessibility. As models grow in size and capability, computational requirements and energy consumption become increasingly significant concerns. Techniques such as sparse attention, mixture-of-experts architectures, quantization, and knowledge distillation will play a central role in making LLMs more efficient, allowing smaller models to perform comparably to massive ones without prohibitive costs. This trend will democratize access to powerful language models for research, industry, and emerging economies, reducing reliance on extremely large proprietary models.

Enhanced reasoning and factual accuracy will also be a focus of future LLM development. Current models, including GPT and T5, can hallucinate or produce plausible yet incorrect information. Integrating external knowledge bases, retrieval-augmented generation, and structured reasoning frameworks may improve the reliability and

trustworthiness of outputs, particularly in applications like scientific research, legal analysis, and medical diagnostics. Additionally, advances in long-context modeling will enable models to handle entire documents, books, or multi-document reasoning tasks more effectively, bridging the gap between short-sequence NLP and extended comprehension tasks.

Ethical alignment and bias mitigation are expected to receive increasing attention. As LLMs are deployed in sensitive contexts, it becomes critical to ensure fairness, inclusivity, and avoidance of harmful outputs. Techniques such as reinforcement learning from human feedback (RLHF), prompt engineering, and adversarial training are likely to evolve to enforce alignment with human values, reduce discriminatory behavior, and ensure safer interactions. Transparency in training data, model interpretability, and robust evaluation frameworks will be central to responsible AI development.

Additionally, hybrid and modular architectures may emerge as a dominant paradigm. Rather than building ever-larger monolithic models, future systems might combine specialized modules for reasoning, language understanding, memory, and multimodal integration. Such modular LLMs could dynamically activate relevant components depending on the task, providing both flexibility and efficiency. This aligns with trends already seen in mixture-of-experts models and adaptive computation techniques.

Finally, the future of LLMs will be shaped by domain specialization and personalization. While general-purpose models like GPT, T5, and LLaMA offer broad capabilities, there is growing demand for models fine-tuned to specific industries, professions, or even individual users. Personalized LLMs could adapt to user preferences, writing styles, or domain-specific knowledge, improving both relevance and usability while maintaining alignment with ethical standards.

In summary, the next generation of language models is likely to emphasize multimodality, efficiency, factual reliability, ethical alignment, modularity, and personalization. While the foundational

architectures of BERT, GPT, T5, and LLaMA remain highly influential, future developments will blend scale with sophistication, enabling models to perform more complex, context-aware, and safe reasoning across a wider spectrum of tasks. The combination of technical innovation, societal responsibility, and accessibility will define the evolution of LLMs, transforming how humans interact with artificial intelligence and leveraging language models as collaborative tools for learning, creativity, and problem-solving.

Large language models (LLMs) have the potential to significantly transform the defense and military sector by enhancing intelligence analysis, decision-making, operational efficiency, and cybersecurity, while also introducing both opportunities and risks that must be carefully managed. The capabilities of models such as BERT, GPT, T5, and LLaMA – ranging from natural language understanding to generative reasoning – can be leveraged to improve situational awareness, strategic planning, and rapid response in complex defense scenarios.

One of the most immediate applications is in intelligence collection and analysis. LLMs can process massive volumes of textual data – including news reports, social media, intercepted communications, and technical documents – far faster than human analysts. By summarizing information, identifying patterns, detecting anomalies, and extracting entities and relationships, these models can provide military analysts with actionable insights in near real-time. For example, BERT-based models excel at information extraction and question-answering tasks, enabling rapid synthesis of structured intelligence from unstructured sources. LLaMA and GPT models, with their generative capabilities, can further support scenario simulation and predictive analysis, generating possible outcomes or strategic recommendations based on textual intelligence.

In decision support and planning, LLMs can assist commanders and analysts by providing natural language briefings, summarizing mission-critical information, and offering scenario-based simulations. Generative models like GPT and T5 can propose multiple operational

options or summarize complex reports into concise actionable items, reducing cognitive load and enabling faster decision-making under pressure. Multimodal LLaMA models could integrate text with satellite imagery, sensor reports, or reconnaissance data, offering richer situational awareness.

LLMs also have potential applications in defense research and development. They can accelerate technical documentation analysis, automate literature reviews, and support the generation of technical proposals or mission-specific plans. By assisting in coding and design tasks, models like GPT-4 can streamline software development for defense applications, such as autonomous systems, simulation tools, or communications protocols. In addition, LLMs can support training and simulation for personnel, generating realistic dialogue scenarios, threat simulations, or educational materials that adapt dynamically to trainee performance.

Cybersecurity and information operations represent another area of impact. LLMs can be used to detect phishing attempts, malware descriptions, or adversarial propaganda in multiple languages, providing automated monitoring and alerting. They can also assist in threat intelligence by summarizing cyber threat reports and predicting potential vulnerabilities. However, the dual-use nature of these models presents risks: the same capabilities could be exploited to generate sophisticated disinformation campaigns, automated phishing messages, or other malicious content, necessitating careful governance and ethical safeguards.

Despite their promise, the integration of LLMs into defense systems comes with critical limitations and challenges. These include susceptibility to errors and hallucinations, biases inherited from training data, and vulnerability to adversarial manipulation. Overreliance on generative outputs without verification could lead to misinformed decisions in high-stakes contexts. Additionally, the computational requirements for training, fine-tuning, and deploying large models may pose logistical challenges for field operations,

particularly in secure or resource-constrained environments. Ensuring security, robustness, and explainability of LLMs in defense applications is therefore paramount.

Looking forward, the future impact of LLMs on the defense industry is likely to grow in strategic and operational domains, where rapid information processing, predictive modeling, and adaptive decision support are critical. By integrating LLMs into intelligence analysis pipelines, simulation frameworks, and cybersecurity operations, defense organizations can achieve unprecedented efficiency and responsiveness. At the same time, the dual-use nature of these technologies underscores the need for strict governance, ethical oversight, and international coordination to prevent misuse and mitigate risks associated with autonomous or semi-autonomous AI in military applications.

In summary, LLMs like BERT, GPT, T5, and LLaMA offer significant opportunities to enhance the defense industry through faster intelligence processing, improved decision support, simulation, research acceleration, and cybersecurity. Their influence will likely grow as multimodal capabilities, alignment, and efficiency improve, enabling more adaptive, informed, and agile military operations. However, responsible deployment and careful management of their limitations, biases, and dual-use potential remain essential to ensuring that these technologies provide strategic advantage without unintended consequences.

Large language models (LLMs) such as BERT, GPT, T5, and LLaMA have the potential to significantly support Ukraine's defense efforts in the ongoing conflict with Russia by enhancing intelligence analysis, situational awareness, decision-making, cybersecurity, and strategic communications. The scale, speed, and language understanding capabilities of these models can provide operational advantages in a rapidly evolving and information-intensive battlefield.

One of the most immediate applications is in real-time intelligence gathering and analysis. LLMs can process vast quantities

of information from multiple sources, including news reports, social media, open-source intelligence (OSINT), and intercepted communications, to identify patterns, track troop movements, detect misinformation, and extract actionable insights. BERT-based models are particularly effective in information extraction and classification tasks, enabling analysts to filter relevant intelligence and prioritize critical information. Generative models such as GPT or LLaMA can further support predictive analyses by generating scenarios, assessing potential outcomes of military operations, or simulating enemy responses based on textual intelligence and historical data.

Decision support and operational planning is another critical area. LLMs can help Ukrainian commanders synthesize complex reports, provide concise summaries of battlefield conditions, and propose alternative courses of action in natural language. This could reduce the cognitive burden on decision-makers during high-pressure situations. The ability of multimodal models like LLaMA to combine textual intelligence with imagery, such as satellite or drone footage, could enhance situational awareness by providing integrated insights from multiple sources in real time.

LLMs also offer support for cybersecurity and counter-information operations, which are essential in the context of hybrid warfare. They can assist in detecting disinformation campaigns, monitoring social media for adversarial propaganda, identifying potential phishing or malware threats, and automating the classification of cyber incidents. This capability is particularly relevant in the Ukraine – Russia conflict, where information warfare has been a central component of the strategic landscape. Generative models could also be used to craft accurate, rapid, and targeted counter-messaging to maintain morale and correct misinformation among civilian and military audiences.

Training and simulation represent additional areas where LLMs can be valuable. Ukrainian forces could leverage these models to develop realistic scenario-based training exercises, generate dynamic operational simulations, and improve tactical decision-making through

adaptive feedback. LLMs can also support research and technical development by rapidly analyzing open-source military literature, technical manuals, or logistical data, helping to optimize strategies, equipment deployment, and resource allocation.

Despite these advantages, there are significant limitations and risks in applying LLMs in the Ukraine – Russia war. Models can generate inaccurate or misleading information if not carefully validated, which could have critical consequences in combat scenarios. They also inherit biases from training data, which could affect intelligence interpretation. Computational requirements for training and deploying large models remain substantial, potentially limiting field-level use without centralized processing infrastructure. Furthermore, adversaries could exploit similar technologies to produce misinformation, automated propaganda, or cyberattacks, creating a dual-use dilemma that must be carefully managed.

In the near term, the most feasible applications for Ukraine may involve intelligence analysis, information verification, decision support, and counter-disinformation strategies, where LLMs can operate as force multipliers for human analysts. Over time, as models become more efficient, multimodal, and better aligned with operational needs, their integration could extend to advanced simulation, predictive modeling, and real-time battlefield support, offering strategic advantages while maintaining safety and control.

In conclusion, LLMs like BERT, GPT, T5, and LLaMA hold significant potential to strengthen Ukraine's defense capabilities in the current conflict by accelerating intelligence processing, enhancing decision-making, supporting cybersecurity, and countering information warfare. Responsible deployment, human oversight, and robust validation are essential to ensure these tools provide reliable support without unintended consequences. When combined with traditional military intelligence and operational expertise, LLMs could play a crucial role in improving situational awareness, operational efficiency, and resilience in the face of hybrid and conventional threats.

The evolution of warfare is increasingly intertwined with advanced information technologies, artificial intelligence, and autonomous systems. In this context, large language models (LLMs) like BERT, GPT, T5, and LLaMA are likely to play a transformative role in future conflicts, influencing decision-making, intelligence, cybersecurity, and information operations. Unlike conventional military technologies that primarily operate in the physical domain, LLMs function in the cognitive and information space, offering the ability to process, analyze, and generate human language at unprecedented speed and scale. This opens new avenues for strategic advantage, rapid decision-making, and operational efficiency, while also raising ethical, security, and operational challenges.

In future wars, LLMs are expected to become critical in intelligence, surveillance, and reconnaissance (ISR). These models can process massive amounts of textual and multimodal data – ranging from social media, news feeds, and diplomatic communications to sensor logs, satellite imagery captions, and technical reports – extracting actionable intelligence in near real-time. Generative capabilities could enable predictive modeling of adversary behavior, simulating potential courses of action, and identifying vulnerabilities in complex operational environments. By automating the initial stages of analysis, LLMs can significantly reduce human cognitive load and accelerate the intelligence cycle, allowing commanders to respond faster to evolving threats.

Another key application is in decision support and command systems. Future conflicts are expected to require rapid interpretation of dynamic situations and coordination across dispersed units and domains. LLMs can provide concise natural-language summaries of battlefield conditions, generate scenario-based recommendations, and assist in planning operations that account for multiple contingencies. Their ability to integrate multimodal data – combining textual intelligence with geospatial information, satellite imagery, or sensor readings – enhances situational awareness and improves the accuracy of strategic and tactical decisions.

Cyber and information warfare will remain a defining aspect of future conflicts, and LLMs are likely to play a dual role. On the defensive side, they can detect disinformation campaigns, identify malicious communications, and support rapid counter-propaganda measures. On the offensive side, adversaries could exploit LLMs to generate sophisticated misinformation, automated phishing campaigns, or social engineering attacks at scale. This dual-use nature underscores the need for ethical frameworks, governance, and robust security protocols in the deployment of LLMs in military contexts.

The integration of LLMs with autonomous and semi-autonomous systems represents another frontier. Future warfighters may leverage LLMs to control or coordinate robotic systems, drones, and logistics operations through natural-language interfaces, reducing operational friction and enabling more agile responses. Instruction-following and multimodal reasoning capabilities in models such as GPT-4 and LLaMA could allow human operators to issue high-level directives that are interpreted and executed by AI-driven platforms in real-time, bridging the gap between strategy and tactical execution.

Despite their potential, LLMs in future warfare face significant limitations and risks. Models may hallucinate information, introduce biases, or misinterpret nuanced intelligence, which could lead to critical errors in high-stakes scenarios. Adversarial actors may attempt to manipulate inputs to mislead AI systems, and the dependence on high-performance computational infrastructure could limit operational flexibility in austere environments. Moreover, the ethical and legal implications of using AI-driven decision-making in lethal or coercive operations require careful consideration, including issues of accountability, transparency, and human oversight.

In conclusion, LLMs are poised to become an essential component of future warfare, particularly in the information and cognitive domains, enhancing intelligence analysis, decision support, cybersecurity, and operational coordination. Their integration into military systems could accelerate decision cycles, improve situational

awareness, and enable adaptive strategies in increasingly complex conflict environments. However, realizing these benefits will require addressing challenges related to reliability, alignment, security, and ethical deployment. As future conflicts continue to blend conventional, cyber, and information operations, LLMs are likely to play a central role in shaping both the conduct and strategic outcomes of warfare, transforming the landscape from purely physical confrontation to a highly integrated cognitive and informational battlespace.

The ongoing Ukraine – Russia war provides a real-world context for understanding how large language models (LLMs) could shape future conflicts, particularly in hybrid and information-centric warfare. The war has highlighted the critical role of rapid intelligence processing, real-time situational awareness, information operations, and cyber capabilities – all areas where LLMs can act as force multipliers. Ukraine's defense efforts illustrate both the opportunities and challenges associated with integrating AI-powered language technologies into military operations.

In intelligence and reconnaissance, LLMs could be employed to process vast amounts of open-source intelligence (OSINT), social media updates, news reports, satellite imagery annotations, and intercepted communications. By automatically extracting entities, classifying threats, and summarizing operational developments, these models can help Ukrainian analysts maintain a comprehensive situational picture in near real time. Generative models, like GPT or LLaMA, could further support predictive analytics, simulating potential enemy maneuvers or logistical bottlenecks and providing probabilistic assessments to guide operational decisions.

Decision support represents another area of potential impact. LLMs can condense complex battlefield reports, generate summaries of dynamic front-line situations, and propose multiple response options, allowing commanders to make faster and more informed choices. Multimodal models that integrate textual intelligence with imagery from drones or satellites could enhance operational planning,

providing a unified understanding of evolving tactical conditions. This ability to combine multiple information streams mirrors the requirements of future conflicts, where speed and comprehension are decisive factors.

Cybersecurity and counter-information operations have been particularly prominent in the Ukraine – Russia war, reflecting the hybrid nature of modern conflicts. LLMs could automate the detection of misinformation campaigns, identify malicious content, and support counter-propaganda efforts aimed at maintaining civilian morale and accurate information dissemination. At the same time, they must be deployed carefully, as adversaries may exploit similar technologies to conduct sophisticated disinformation campaigns or cyberattacks at scale.

LLMs could also enhance training, simulations, and operational planning. By generating realistic scenarios, interactive simulations, and dynamic threat assessments, these models can accelerate the preparation of military personnel, improve tactical decision-making, and optimize resource allocation. Additionally, they can assist in technical research, such as analyzing engineering documents, operational manuals, or logistics plans, to streamline support and innovation in the field.

Despite these advantages, Ukraine's experience underscores the limitations and risks of relying on LLMs in conflict. Models are prone to hallucinations or misinterpretation, which could lead to erroneous conclusions in high-stakes environments. Biases in training data may affect intelligence assessments, and the computational demands of large models pose challenges for deployment in forward-operating or resource-constrained conditions. Human oversight, verification of outputs, and robust validation processes remain essential to prevent operational errors or the misuse of AI-generated information.

Ukraine's conflict demonstrates that future wars will increasingly be hybrid, combining kinetic operations with cyber, information, and

cognitive domains. LLMs, by enabling rapid synthesis of diverse information sources, scenario modeling, and automated decision support, could play a central role in this multidimensional battlespace. The experience also highlights the importance of resilience, adaptability, and ethical deployment: effective use of LLMs requires not only technical sophistication but also strategic integration with human decision-making and broader military capabilities.

In summary, the Ukraine – Russia war illustrates a microcosm of the future battlefield, where LLMs have the potential to enhance intelligence, decision-making, cybersecurity, and information operations. By acting as cognitive force multipliers, these models can improve situational awareness, speed up decision cycles, and support complex operational planning. At the same time, they emphasize the continuing need for human judgment, rigorous verification, and careful governance to ensure that the benefits of AI technologies are realized without introducing undue risks in high-stakes military contexts.

Conclusions. Large language models (LLMs) such as BERT, GPT, T5, and LLaMA represent a transformative leap in the field of natural language processing. Each of these models embodies distinct architectural principles that define its strengths, limitations, and potential applications. BERT, with its bidirectional encoder architecture, excels at understanding and extracting meaning from text, providing state-of-the-art performance in comprehension-focused tasks. GPT, a decoder-only autoregressive model, demonstrates exceptional generative capabilities, enabling coherent text production, scenario simulation, and conversational AI. T5 unifies diverse NLP tasks under a text-to-text encoder-decoder framework, allowing seamless transition between comprehension and generation, while LLaMA emphasizes efficient scaling and research accessibility, combining strong generative performance with versatility for fine-tuning across tasks.

The limitations of these models highlight inherent trade-offs. BERT cannot generate text and is constrained by fixed input lengths; GPT is prone to hallucinations and factual inaccuracies and requires

substantial computational resources; T5, although highly versatile, is sensitive to prompt formulation and resource-intensive; LLaMA, while efficient and scalable, still inherits biases, and its multimodal capabilities remain limited in early implementations. Recognizing these constraints is essential for responsible deployment, especially in high-stakes applications such as defense, intelligence, and decision-making.

The future development of LLMs is likely to focus on several key areas: multimodal integration, which combines text with images, video, and sensor data; efficiency improvements through sparsity, quantization, and mixture-of-experts architectures; enhanced factual accuracy via retrieval-augmented generation and structured reasoning; ethical alignment and bias mitigation through reinforcement learning from human feedback and governance frameworks; modular architectures that allow adaptive computation depending on task requirements; and domain specialization and personalization for more relevant and effective applications. These developments will enable LLMs to become more versatile, reliable, and accessible, while addressing the growing societal and operational expectations for AI systems.

In the context of defense and national security, LLMs offer transformative capabilities. They can accelerate intelligence analysis, support rapid decision-making, enhance situational awareness, improve cybersecurity defenses, and assist in training and simulation. Models such as BERT can extract structured insights from massive textual datasets, GPT and T5 can generate scenario-based predictions and actionable intelligence, and LLaMA can provide efficient, research-oriented generative capabilities. Ukraine's experience in the ongoing conflict with Russia illustrates the practical potential of these technologies in hybrid warfare, where real-time intelligence, information verification, counter-disinformation, and decision support are critical. LLMs can act as cognitive force multipliers, processing large-scale data streams, summarizing complex information, and providing predictive insights that support operational effectiveness.

Looking toward future conflicts, LLMs are expected to play a central role in the cognitive and information domains of warfare. They will enable rapid situational assessment, predictive modeling, autonomous decision support, and integration with multimodal intelligence and autonomous systems. However, their deployment comes with inherent risks, including hallucinations, biases, adversarial manipulation, and reliance on high-performance computational infrastructure. Ethical, legal, and operational governance will therefore be essential to ensure responsible use, reliability, and safety.

In conclusion, LLMs represent both an extraordinary technological opportunity and a set of complex challenges. Their architectures, capabilities, and limitations define their suitability for diverse tasks, from general-purpose NLP applications to high-stakes military operations. Future research and development will likely focus on scaling efficiency, integrating multimodal intelligence, improving reasoning and factual accuracy, and aligning these models with human values. When deployed responsibly, LLMs have the potential to reshape the landscape of human-computer interaction, strategic decision-making, and the conduct of future warfare, providing unprecedented advantages in processing, understanding, and generating information. The ongoing Ukraine – Russia conflict serves as a contemporary example of how these models can enhance operational effectiveness, while also highlighting the importance of oversight, validation, and ethical deployment in real-world scenarios.

References

1. Baevski, M. Auli, Adaptive input representations for neural language modeling, arXiv preprint arXiv:1809.10853 (2018).

2. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. Lima: Less is more for alignment, arXiv preprint arXiv:2305.11206 (2023).

3. Schuurmans, D. Memory augmented large language models are computationally universal, arXiv preprint arXiv:2301.04589 (2023)

4. DeepSeek-AI, Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Deng, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Yang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Chen, J., Yuan, J., Qiu, J., Song, J., Dong, K., Gao, K., Guan, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Pan, R., Xu, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Zheng, S., Wang, T., Pei, T., Yuan, T., Sun, T., Xiao, W. L., Zeng, W., An, W., Liu, W., Liang, W., Gao, W., Zhang, W., X. Q. Li, Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Chen, X., Nie, X., Sun, X., Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, CoRR abs/2405.04434 (2024).

5. Zhang, H., Dang, M., Peng, N., and den Broeck, G. V. “Tractable Control for Autoregressive Language Generation.” arXiv, Apr. 17, 2023. [Online]. Available: http://arxiv.org/ abs/2304.07438

6. Chen, L., Zaharia, M., and Zou, J. “FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance.” arXiv, May 09, 2023. [Online]. Available: http:// arxiv.org/abs/2305.05176

7. Xue, L. et al., “ULIP-2: Towards Scalable Multimodal Pre-training For 3D Understanding.” arXiv, May 14, 2023. [Online]. Available: http://arxiv.org/abs/2305.08275

8. Kwon, M., Hu, H., Myers, V., Karamcheti, S., Dragan, A., and Sadigh, D. “Toward Grounded Social Reasoning.” arXiv, Jun. 14, 2023. [Online]. Available: http://arxiv.org/ abs/2306.08651

9. Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., Gao, J. Chameleon: Plug-and-play compositional reasoning with large language models, arXiv preprint arXiv:2304.09842 (2023).

10. Sennrich, R., Haddow, B., Birch, A. Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1715–1725.

11. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., Stojnic, R. Galactica: A large language model for science, arXiv preprint arXiv:2211.09085 (2022).

12. Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., et al., Opt-iml: Scaling language model instruction meta learning through the lens of generalization, arXiv preprint arXiv:2212.12017 (2022).

13. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A. Progprompt: Generating situated robot task plans using large language models, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 11523–11530. 20, 33.

14. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness, Advances in Neural Information Processing Systems 35 (2022) 16344–16359.

15. Naseem, U., Razzak, I., Khan, S. K., Prasad, M. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models, Transactions on Asian and LowResource Language Information Processing 20(5) (2021).

16. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface, arXiv preprint arXiv:2303.17580 (2023).

17. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context, arXiv preprint arXiv:1901.02860 (2019).

18. Gekhman, Z., Herzig, J., Aharoni, R., Elkind, C., and Szpektor, I. "TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models." arXiv, May 18, 2023. [Online]. Available: http://arxiv.org/abs/2305.11171

19. Luo, Z. et al. "Augmented Large Language Models with Parametric Knowledge Guiding." arXiv, May 08, 2023. [Online]. Available: http://arxiv.org/abs/2305.04757

20. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. The rise and potential of large language model based agents: A survey, arXiv preprint arXiv:2309.07864 (2023).

3.2. OPEN-SOURCE DATA, SOCIAL GRAPHS, AND BOT NETWORKS: DETECTING INFLUENCE CAMPAIGNS IN REAL TIME

Introduction. In today's hyperconnected digital ecosystem, social media platforms have become the primary arenas for communication, information exchange, and public discourse. While these platforms enable unprecedented con – nectivity and democratize the sharing of ideas, they also present new avenues for manipulation, misinformation, and orchestrated influence campaigns. Actors ranging from state-sponsored entities to ideological groups and commercial interests can exploit the open nature of social networks to sway public opinion, amplify narratives, and distort online discourse. Detecting and countering such influence operations has thus emerged as a critical challenge for cybersecurity experts, data scientists, and policy makers alike.

A key approach to understanding and mitigating these threats lies in the analysis of open-source data. Unlike proprietary datasets, open-source information – including publicly available social media posts, user profiles, network interactions, and metadata – offers a wealth of insights that can be harnessed without breaching privacy or legal constraints. By systematically aggregating and analyzing this data, researchers can identify unusual patterns of activity, map relationships between accounts, and detect coordinated behaviors indicative of manipulation.

Social graph analysis forms the backbone of this effort. Social graphs represent the complex web of connections among users, allowing analysts to trace information flow, uncover communities, and identify central nodes that exert disproportionate influence. By examining metrics such as centrality, clustering, and network density, it becomes possible to distinguish organic interactions from artificial or coordinated campaigns. When combined with temporal analysis, these methods can reveal sudden bursts of activity,

synchronized posting, or content amplification patterns that are characteristic of bot-driven operations.

Bot networks, in particular, represent one of the most challenging aspects of influence campaigns. Automated accounts can amplify content at scale, simulate human behavior, and create the illusion of widespread consensus. Sophisticated bots can engage in targeted interactions, mimic sentiment, and even evolve over time to avoid detection. Understanding the structure and behavior of these networks is therefore essential for real-time detection of influence operations. Machine learning algorithms, anomaly detection techniques, and graph-based methods provide powerful tools for identifying bots and their orchestrators within the broader social graph.

Real-time detection is critical because influence campaigns are most effective when they can shape narratives quickly and exploit viral dynamics. Delays in identifying coordinated manipulation allow false or misleading information to spread, potentially affecting elections, public health responses, and societal stability. Integrating open-source intelligence, social graph analytics, and bot detection techniques into automated monitoring systems enables timely intervention, allowing platforms, governments, and civil society organizations to respond rapidly to emerging threats.

In summary, the intersection of open-source data, social network analysis, and bot detection offers a promising avenue for addressing the growing challenge of influence campaigns. By leveraging these methodologies, researchers and practitioners can not only map and understand digital manipulation but also develop actionable strategies to mitigate its impact in real time. This multidisciplinary approach, bridging computer science, social science, and cybersecurity, is essential for preserving the integrity of online information ecosystems in an era of increasingly sophisticated digital threats. The importance of open-source data in this domain cannot be overstated. Unlike private or proprietary datasets, open-source information is continuously updated and widely accessible, making

it a practical foundation for real-time monitoring. Analysts can track trends, detect anomalies, and uncover coordinated behaviors without infringing on user privacy, relying on posts, likes, shares, retweets, hashtags, and other publicly available interactions. This democratization of data enables not only academic research but also proactive threat detection by governments, non-governmental organizations, and platform operators.

Social graphs, as representations of relationships and interactions among users, provide a powerful lens to visualize and quantify the spread of information. These networks are not merely static maps; they evolve dynamically, reflecting the flow of narratives, the emergence of clusters, and the influence of individual nodes. By applying graph-theoretic measures – such as betweenness centrality, modularity, and network assortativity – researchers can identify influential actors, hidden communities, and patterns suggestive of manipulation. Furthermore, temporal network analysis allows the detection of coordinated bursts of activity, which often signal orchestrated campaigns rather than organic discussions.

Bot networks are particularly insidious because they can mimic authentic human engagement at scale. Automated accounts can generate and amplify content across multiple platforms, creating an illusion of consensus or popularity. Advanced bots are capable of adaptive behavior, learning from interactions to evade simple detection heuristics. They may retweet selectively, engage in conversations strategically, or synchronize activity across multiple accounts. Understanding and modeling these networks require sophisticated computational approaches, including machine learning classifiers, clustering algorithms, and anomaly detection techniques, all integrated within the broader framework of social graph analysis.

Real-time detection of influence campaigns is not merely a technical challenge but a societal necessity. Misinformation and disinformation can spread at unprecedented speeds, influencing political outcomes, public health decisions, and social cohesion.

Early detection allows for timely countermeasures, such as content moderation, user alerts, and network disruption, minimizing the potential damage. By combining open-source intelligence, social graph insights, and automated bot detection, it is possible to create a comprehensive system capable of identifying emerging threats before they reach critical mass.

Ultimately, the study of influence campaigns through open-source data, social networks, and bot analysis represents a convergence of multiple disciplines: computer science, network theory, cybersecurity, and social psychology. It highlights the evolving nature of digital information ecosystems, the sophistication of modern manipulation techniques, and the urgent need for real-time monitoring and intervention. By advancing methods for detection and analysis, researchers contribute not only to the security of online platforms but also to the broader goal of safeguarding informed and resilient societies.

Presentation of the main research material. The main research material for this study encompasses three interrelated domains: open-source data, social graph structures, and bot networks. Each of these components provides a unique perspective on the mechanisms underlying influence campaigns and forms the foundation for detection and analysis in real time. By integrating these data sources and analytical methods, it is possible to construct a comprehensive framework for understanding and mitigating digital manipulation.

Open-Source Data: The research utilizes publicly available social media content, including posts, comments, likes, shares, hashtags, user metadata, and timestamps. This dataset provides insight into both the content of messages and the patterns of interaction among users. By focusing on open-source information, the study ensures transparency, reproducibility, and compliance with privacy regulations while enabling large-scale monitoring of potential influence operations.

Social Graphs: Social graphs serve as the structural backbone for analyzing interactions and information flow. Nodes represent individual accounts or users, while edges represent interactions such

as retweets, replies, or shared content. Network analysis techniques – including centrality measures, community detection algorithms, and temporal network evaluation – allow the identification of influential nodes, tightly-knit clusters, and anomalies that may indicate coordinated activity. These graphs are essential for mapping both overt and hidden relationships between actors within digital ecosystems.

Bot Networks: Automated accounts, or bots, are a central focus of this research due to their role in amplifying and coordinating influence campaigns. The material includes datasets on account activity patterns, posting frequency, content similarity, and engagement metrics. Analytical methods such as machine learning classification, anomaly detection, and network clustering are applied to identify bot accounts, distinguish them from human users, and reveal the structure of coordinated networks. The study emphasizes not only the detection of individual bots but also the identification of larger orchestrated networks that drive influence campaigns.

Open-source data serves as the foundation for analyzing influence campaigns in digital environments. Unlike proprietary datasets, open-source data is publicly accessible and continuously updated, providing researchers with a rich resource for studying social interactions, content dissemination, and behavioral patterns across platforms. The scope of open-source data includes textual content, multimedia posts, metadata, network interactions, and publicly available user profiles. Collectively, these elements offer a comprehensive view of the dynamics within online communities.

Data Types and Sources: The primary types of open-source data used in this research include posts, comments, shares, retweets, likes, hashtags, mentions, and timestamps. Social media platforms such as Twitter, Reddit, Facebook (public pages and groups), Instagram (public accounts), and other online forums provide this data. Additionally, web scraping of public blogs, news outlets, and discussion boards complements social media data, enabling a broader perspective on information propagation and discourse patterns.

Advantages of Open-Source Data: One of the main advantages of open-source data is its accessibility and legal transparency. Researchers can collect, analyze, and share findings without infringing on privacy regulations or platform restrictions, which is critical for reproducibility and transparency in academic studies. Moreover, open-source datasets reflect real-time activity, enabling the detection of emerging trends, viral content, and coordinated campaigns as they unfold.

Challenges and Limitations: While open-source data offers significant benefits, it also presents challenges. Data can be noisy, incomplete, or biased due to platform-specific features, algorithmic content curation, and self-selection of users who choose to post publicly. Additionally, the sheer volume and velocity of data require robust collection, storage, and processing mechanisms. Researchers must employ techniques for filtering irrelevant content, normalizing metadata, and managing large-scale datasets efficiently.

Applications in Influence Campaign Detection: Open-source data enables multiple layers of analysis. Content-based analysis allows identification of recurring narratives, sentiment trends, and potentially misleading information. Interaction-based analysis – such as tracking likes, shares, or replies – helps reveal patterns of engagement that may indicate artificial amplification. When combined with temporal analysis, open-source data can uncover bursts of coordinated activity, suggesting the presence of orchestrated campaigns or bot networks.

In conclusion, open-source data is indispensable for studying influence campaigns. Its accessibility, richness, and real-time nature make it a powerful tool for detecting manipulation and understanding the dynamics of online discourse. By systematically collecting, processing, and analyzing open-source data, researchers can construct the empirical basis for identifying suspicious behaviors and developing strategies to counteract digital influence operations.

Data Collection Techniques: Collecting open-source data for influence campaign research involves multiple approaches. Application Programming Interfaces (APIs) provided by social

media platforms are the most common and reliable method, allowing structured access to posts, comments, likes, retweets, and user metadata. For platforms or websites without public APIs, web scraping and crawling techniques are employed, enabling the extraction of publicly visible content while respecting legal and ethical boundaries. Additionally, third-party aggregators and archives can provide historical datasets, which are crucial for longitudinal studies of influence campaigns and behavioral trends.

Data Preprocessing: Raw open-source data often requires extensive preprocessing before analysis. Textual data must be cleaned to remove noise such as spam, advertisements, irrelevant hashtags, or repeated messages. Natural Language Processing (NLP) techniques – including tokenization, lemmatization, and sentiment analysis – are applied to extract meaningful patterns from text. Metadata, such as timestamps, geolocation tags, and user profile attributes, is standardized and normalized to facilitate temporal and spatial analyses. Filtering and deduplication ensure that repetitive or automated content does not skew results.

Analytical Applications: Open-source data enables multiple analytical layers:

1. Content Analysis: NLP and topic modeling can identify dominant themes, narratives, and sentiment trends within public discourse. Repeated or coordinated messaging can indicate attempts to manipulate opinion or propagate misinformation.

2. Temporal Analysis: By tracking posts over time, researchers can detect sudden surges of activity, repeated posting patterns, or synchronization across accounts – typical markers of orchestrated influence campaigns.

3. Interaction Analysis: Public engagement metrics (likes, shares, retweets, comments) allow the identification of highly influential nodes and the mapping of information propagation pathways. Patterns such as disproportionate amplification of specific content may indicate artificial promotion by bot networks or coordinated human actors.

4. Cross-Platform Tracking: Influence campaigns often span multiple platforms. By aggregating open-source data from different social networks, researchers can map the flow of narratives across digital ecosystems, detect coordinated messaging strategies, and identify central actors who bridge communities.

Challenges and Ethical Considerations: While open-source data is legally and publicly accessible, ethical concerns remain. Researchers must ensure anonymization of sensitive personal information, avoid targeted surveillance of individuals, and comply with platform-specific terms of service. Additionally, the data is inherently biased – certain demographics may be overrepresented, and algorithmic amplification on platforms can distort the apparent popularity of content. Analysts must account for these biases to avoid misleading conclusions.

Significance for Real-Time Detection: The real power of open-source data lies in its ability to support real-time or near-real-time monitoring. By continuously collecting, processing, and analyzing public content, researchers can detect emerging influence campaigns, track their evolution, and respond proactively. Integration with automated detection systems, such as anomaly detection algorithms or bot identification tools, enhances the ability to mitigate the impact of digital manipulation before it spreads widely.

Open-source data has played a pivotal role in the evolution and advancement of large language models (LLMs). These models, including GPT-style architectures, rely on massive datasets to learn linguistic patterns, semantic relationships, and contextual reasoning. Open-source data, encompassing publicly available text from websites, forums, social media, academic papers, code repositories, and other digital content, provides the scale, diversity, and richness necessary to train models capable of understanding and generating human-like language.

Scale and Diversity of Training Data: LLMs require vast and varied datasets to generalize effectively across different topics, styles, and domains. Open-source data supplies a rich tapestry of language from multiple sources, cultures, and contexts. For instance, datasets

such as Wikipedia, GitHub repositories, Project Gutenberg texts, and public forums offer a wide range of vocabulary, syntax, and semantic structures. This diversity allows models to learn not only common linguistic patterns but also nuanced idiomatic expressions, technical terminology, and domain-specific knowledge.

Transparency and Reproducibility: Open-source datasets enhance transparency in AI research. They enable independent verification of training procedures, performance benchmarks, and model behavior, fostering a collaborative environment where improvements and innovations can be shared openly. Open-source corpora also support reproducibility, allowing researchers worldwide to replicate experiments and refine models without relying exclusively on proprietary or private datasets, which may be inaccessible or restricted.

Accelerating Innovation: The availability of open-source data has accelerated innovation in natural language processing. Researchers can experiment with novel architectures, pretraining objectives, and fine-tuning techniques using publicly available datasets. Open-source initiatives, such as The Pile, Common Crawl, and various academic corpora, have significantly lowered the barriers to entry, enabling startups, academic institutions, and independent researchers to develop competitive LLMs without massive proprietary datasets.

Bias and Quality Considerations: While open-source data provides scale and diversity, it also introduces challenges. Open-source corpora may contain biases, misinformation, or unbalanced representation of certain groups or topics. LLMs trained on such data can inadvertently learn and reproduce these biases, making careful curation, filtering, and preprocessing essential. Nonetheless, these challenges have driven the development of advanced data-cleaning techniques, alignment methods, and post-training evaluation frameworks, contributing to more robust and ethical LLM deployment.

Real-World Applications: The influence of open-source data on LLMs extends to their real-world utility. Models trained on extensive open-source content can perform a wide range

of tasks, including text summarization, question answering, code generation, sentiment analysis, and content moderation. The breadth of open-source knowledge embedded in these models allows them to understand domain-specific queries, generate contextually accurate responses, and assist in research, education, and decision-making.

In conclusion, open-source data has been a transformative force in the development of LLMs. By providing the scale, diversity, and transparency needed for effective model training, it has democratized access to cutting-edge AI, accelerated innovation, and enhanced the versatility of language models. At the same time, it has highlighted the need for careful data curation, bias mitigation, and ethical considerations, shaping both the capabilities and responsibilities of modern AI systems.

Social graphs are a fundamental tool for analyzing the structure and dynamics of online interactions. They provide a mathematical and visual representation of relationships between users, content, and communities within digital ecosystems. In the context of influence campaigns and digital manipulation, social graphs enable researchers to identify influential actors, uncover coordinated activity, and map the propagation of information across networks.

Definition and Structure: A social graph consists of nodes and edges. Nodes represent individual accounts, users, or entities, while edges represent relationships or interactions between them, such as friendships, follows, mentions, retweets, replies, or shared content. Social graphs can be directed (where relationships have direction, e.g., follower → followed) or undirected (where connections are mutual). The graph structure captures not only direct interactions but also the broader network topology, revealing clusters, communities, and the centrality of specific nodes.

Analytical Metrics: Social graph analysis employs a range of metrics to identify key features and patterns within networks:

• Centrality Measures: Metrics such as degree centrality, betweenness centrality, and eigenvector centrality highlight nodes with

significant influence or control over information flow. High-centrality nodes often correspond to opinion leaders or amplifiers within a network.

• Community Detection: Algorithms such as modularity optimization or spectral clustering identify tightly connected groups of nodes. Coordinated campaigns often appear as dense clusters with frequent internal interactions.

• Network Density and Connectivity: Measures of edge density, clustering coefficient, and network diameter indicate how interconnected a community is, providing insights into the potential for rapid information dissemination.

• Temporal Analysis: Social graphs are dynamic, evolving over time. Temporal network analysis allows the detection of sudden bursts of activity, synchronized posting, or rapid spreading of specific content – hallmarks of orchestrated campaigns.

Applications in Influence Campaign Detection: Social graphs are instrumental in detecting both organic and coordinated manipulation:

• Influence Mapping: By analyzing central nodes and information flow, researchers can identify accounts that disproportionately shape discourse.

• Bot Network Detection: Dense clusters of highly active, similarly behaving nodes often indicate automated accounts. By combining graph topology with behavioral metrics, bot networks can be distinguished from genuine human interactions.

• Propagation Analysis: Tracking how content spreads through the graph reveals amplification strategies, viral campaigns, and the relative influence of individual nodes or groups.

Visualization and Interpretation: Visual representations of social graphs enhance interpretability, allowing researchers to observe clusters, outliers, and structural patterns that may be less obvious from raw data. Graph visualization tools and software, such as Gephi, NetworkX, or Cytoscape, support interactive exploration and facilitate communication of findings to both technical and non-technical audiences.

Challenges and Limitations: While social graph analysis is powerful, it also faces challenges. Real-world social networks are massive, dynamic, and heterogeneous, requiring efficient algorithms and high computational resources. Data sparsity, missing interactions, or platform-imposed limitations can reduce accuracy. Moreover, malicious actors can intentionally obscure their networks to evade detection, necessitating sophisticated analytical techniques.

In summary, social graphs provide a robust framework for understanding the structure and dynamics of online interactions. When combined with open-source data and bot detection methods, they enable comprehensive detection of influence campaigns, offering insights into both the actors involved and the mechanisms by which information spreads in digital ecosystems.

Dynamic and Temporal Social Graphs: Real-world social networks are not static; interactions evolve over time. Temporal social graphs incorporate the dimension of time, allowing researchers to track how connections, influence, and information propagation change dynamically. This temporal perspective is crucial for detecting influence campaigns, as coordinated activities often occur in synchronized bursts, exploiting specific events or trending topics. Temporal analysis can reveal patterns such as repeated content amplification, simultaneous account activation, or rapid formation of dense interaction clusters – all indicators of orchestrated campaigns.

Weighted and Multi-Layer Graphs: Not all connections in a social graph are equal. Weighted graphs assign importance or strength to edges, reflecting the frequency, intensity, or significance of interactions. For example, repeated retweets or frequent mentions indicate stronger influence than a single interaction. Multi-layer graphs extend this idea by representing different types of relationships separately – such as follower connections, content sharing, and direct messages – allowing more nuanced analysis of complex social dynamics.

Community Detection and Structural Analysis: Detecting communities within social graphs is vital for identifying groups

of users that may be coordinating behavior. Algorithms like Louvain modularity, Infomap, or spectral clustering can uncover tightly connected clusters that exhibit high internal interaction and low external connectivity. In influence campaign research, such clusters often correspond to bot networks or coordinated human actors propagating specific narratives. Structural metrics, such as clustering coefficient, network modularity, and assortativity, provide additional insight into network cohesion and potential manipulative activity.

Influence Propagation and Centrality: Social graphs allow modeling how information spreads through a network. Centrality measures (degree, betweenness, closeness, eigenvector) help identify influential nodes that can disproportionately affect information flow. Nodes with high betweenness centrality, for example, often act as bridges between communities, facilitating cross-cluster dissemination of narratives. Monitoring the activity and influence of these nodes can provide early warning signs of emerging campaigns.

Anomaly Detection Using Graphs: Graph-based anomaly detection techniques are critical for spotting suspicious behavior. Sudden spikes in connectivity, unusually dense clusters, or repetitive interaction patterns can indicate automated amplification or coordinated efforts. Machine learning models, such as graph neural networks (GNNs) or community-aware anomaly detectors, are increasingly applied to detect these patterns in large-scale social graphs.

Integration with Other Data Sources: Social graphs are most powerful when combined with open-source data and behavioral analysis. By linking graph structure with content semantics, temporal patterns, and user metadata, researchers can build comprehensive models for detecting influence campaigns. For example, a dense cluster of accounts sharing highly similar content at synchronized times is far more indicative of coordinated activity than isolated interactions alone.

Visualization and Practical Applications: Advanced visualization techniques help interpret complex social graphs. Heatmaps, interactive network diagrams, and time-lapse visualizations allow analysts to spot

clusters, key influencers, and propagation pathways effectively. In practical applications, social graph analysis informs strategies for content moderation, bot mitigation, public awareness campaigns, and real-time threat detection.

Challenges and Limitations: Despite their utility, social graphs face challenges in real-world application. Networks can be extremely large and dynamic, making computation and storage intensive. Data incompleteness, platform restrictions, and deliberate obfuscation by malicious actors can reduce accuracy. Additionally, interpreting complex graph metrics requires expertise to distinguish genuine influence from coincidental network structure.

In summary, advanced social graph analysis provides an indispensable framework for mapping and understanding online interactions. By leveraging temporal, weighted, and multi-layered graph models, researchers can uncover patterns of coordination, detect bot networks, and monitor the spread of influence in real time. When integrated with open-source data and bot detection methods, social graphs offer a robust and scalable approach to understanding and mitigating digital manipulation campaigns.

Bot networks, or automated account clusters, play a central role in orchestrated influence campaigns. These networks consist of accounts controlled by scripts or automated systems rather than human users, designed to amplify specific messages, manipulate public perception, or create the illusion of consensus. Understanding the structure, behavior, and detection of bot networks is essential for identifying and mitigating coordinated online manipulation.

Characteristics of Bot Accounts: Bots exhibit distinctive behavioral patterns that differentiate them from organic human users. Common characteristics include:

- High posting frequency: Bots often post at rates far exceeding typical human activity.
- Synchronized behavior: Multiple accounts may share, retweet, or post identical or highly similar content simultaneously.

• Repetitive messaging: Bots frequently recycle the same messages or hashtags across multiple accounts.

• Limited interaction diversity: While human users engage in varied conversations, bots often interact primarily with a small set of accounts or content sources.

• Profile anomalies: Bots may have incomplete profiles, generic images, or irregular follower-to-following ratios.

Bot Networks and Social Graphs: Bot accounts rarely operate in isolation; they are typically organized into networks that amplify messages and influence target audiences. Social graph analysis is crucial for uncovering these networks, as bots often form dense clusters with high internal connectivity and synchronized activity. By mapping interactions, researchers can identify coordinated behavior that might not be apparent when examining individual accounts. Multi-layer graph models can reveal connections across different platforms or communication channels, exposing complex cross-platform bot networks.

Detection Techniques: Detecting bots and bot networks involves a combination of statistical, graph-based, and machine learning approaches:

• Behavioral Analysis: Metrics such as posting frequency, temporal patterns, and engagement rates help flag suspicious accounts.

• Content Similarity: Natural language processing (NLP) techniques identify repeated or templated messages indicative of automated activity.

• Graph-Based Detection: Network metrics, including cluster density, centrality patterns, and anomalous connectivity, help distinguish coordinated bot networks from organic user communities.

• Machine Learning and AI: Supervised and unsupervised learning models, including random forests, support vector machines, and graph neural networks (GNNs), are increasingly used to classify accounts and detect network-level coordination.

Real-Time Monitoring: One of the most critical aspects of bot detection is the ability to operate in real time. Influence campaigns

often rely on rapid dissemination of messages to achieve viral impact. Automated monitoring systems that continuously collect open-source data, update social graphs, and apply bot detection algorithms can detect suspicious activity as it emerges, enabling timely countermeasures such as content moderation, account suspension, or public alerts.

Challenges in Detection: Bot networks are becoming increasingly sophisticated. Modern bots can mimic human behavior, vary their posting schedules, interact naturally with content, and evolve to avoid detection. These advances require adaptive detection techniques, continuous model retraining, and integration of multi-modal data sources. False positives also pose a challenge; misclassifying genuine users as bots can undermine trust and raise ethical concerns.

Applications and Implications: Understanding bot networks has broad applications:

- Influence Campaign Mitigation: Detecting and disrupting bot networks can reduce the spread of misinformation and prevent coordinated manipulation of public opinion.
- Policy and Governance: Insights from bot network analysis inform platform policies, governmental regulations, and public awareness strategies.
- Research and Social Science: Studying bot networks provides a window into modern information warfare, digital sociology, and the dynamics of online communities.

In conclusion, bot networks represent a critical component of contemporary influence campaigns. Through a combination of behavioral analysis, social graph techniques, and machine learning, researchers can identify and mitigate automated amplification and coordinated activity. Integrating bot network detection with open-source data and social graph analysis provides a comprehensive framework for understanding, tracking, and countering digital manipulation in real time.

Detecting influence campaigns in real time requires a holistic approach that combines multiple data sources and analytical

frameworks. By integrating open-source data, social graph analysis, and bot network detection, researchers and practitioners can construct a comprehensive system capable of identifying, tracking, and mitigating coordinated manipulation as it unfolds.

1. Data Collection and Preprocessing:Open-source data serves as the primary input for the system. Publicly available posts, comments, retweets, likes, hashtags, and user metadata are continuously collected from multiple platforms. Preprocessing techniques – including noise filtering, text normalization, and metadata standardization – prepare the data for subsequent analysis. Temporal tagging and cross-platform aggregation allow the system to detect synchronous activity and patterns of content propagation that may indicate influence operations.

2. Social Graph Construction and Analysis:Collected data is transformed into social graphs, representing nodes (users or accounts) and edges (interactions such as mentions, shares, or replies). Graph-based analysis enables the identification of key influencers, tightly connected communities, and unusual structural patterns. Temporal and weighted graphs enhance detection capabilities by capturing the timing, frequency, and strength of interactions. Advanced metrics such as centrality, clustering coefficient, and modularity help highlight nodes and clusters that may be driving coordinated campaigns.

3. Bot Network Identification:Automated accounts are often central to influence campaigns. Behavioral patterns, content similarity, and network structure are analyzed to detect individual bots and coordinated bot networks. Graph-based anomaly detection, combined with machine learning models, enables the system to differentiate between organic human activity and orchestrated amplification. Real-time monitoring ensures that new bots or emerging clusters are quickly identified before they can significantly impact public discourse.

4. Integrated Detection Pipeline: By combining these components, the system operates as an integrated detection pipeline:

• Open-source data provides the raw information and temporal context.

• Social graphs reveal the structural and relational dynamics of information spread.

• Bot detection isolates automated actors and coordinated networks.

Together, these layers allow the system to detect influence campaigns at both the micro level (individual bot accounts or coordinated messages) and macro level (overall network manipulation and narrative propagation).

5. Real-Time Alerts and Mitigation:Integration enables proactive responses. Detected campaigns can trigger alerts for analysts, content moderation actions, or automated mitigation strategies. Temporal and behavioral analysis ensures that emerging campaigns are flagged quickly, reducing the likelihood of widespread misinformation dissemination. Continuous feedback from detected events also allows the system to improve detection accuracy over time, adapting to evolving tactics used by malicious actors.

6. Benefits and Applications: The integrated approach offers multiple advantages:

• Timeliness: Detects campaigns as they develop, rather than post-factum.

• Scalability: Handles large volumes of open-source data across platforms.

• Comprehensiveness: Captures structural, behavioral, and content-based indicators.

• Actionability: Supports intervention strategies for platforms, policymakers, and researchers.

The integration of open-source data, social graph analysis, and bot network detection forms a robust framework for real-time influence campaign detection. By leveraging these complementary methods, it is possible to uncover coordinated manipulation, map the flow of information, and identify actors driving narratives. This

holistic approach is essential for maintaining the integrity of online discourse, protecting public opinion, and responding proactively to the increasingly sophisticated tactics of digital influence operations.

The integration of open-source data, social graph analysis, and bot network detection has significant implications for information security. In today's digital ecosystem, where misinformation, disinformation, and coordinated influence campaigns can spread rapidly, these methodologies provide critical tools for safeguarding the integrity, availability, and trustworthiness of information.

1. Early Detection of Threats: Open-source data enables continuous monitoring of public digital spaces, allowing security analysts to detect emerging threats before they escalate. By tracking content, user behavior, and network activity in real time, potential influence campaigns – whether politically motivated, ideologically driven, or commercially orchestrated – can be identified early, reducing their impact on public discourse and decision-making.

2. Mapping and Mitigating Coordinated Campaigns: Social graph analysis exposes the structural dynamics of information flow. By identifying highly connected nodes, dense clusters, and cross-community bridges, analysts can pinpoint actors who have the potential to amplify false narratives or manipulate opinions. Detecting these patterns allows for targeted mitigation strategies, such as disrupting bot networks, limiting the reach of manipulated content, or alerting users to potential misinformation.

3. Reducing the Impact of Automated Threats: Bot networks pose a substantial risk to information security because they can amplify disinformation at scale and create artificial consensus. Detection of bot behavior – through graph analysis, behavioral monitoring, and machine learning – enables security systems to neutralize these automated threats, preserving the authenticity of online discourse. Removing or isolating bot networks reduces the likelihood that malicious actors can manipulate perception, incite unrest, or exploit public sentiment.

4. Enhancing Trust and Resilience: The combined use of open-source data, social graphs, and bot detection contributes to overall trust in digital platforms. By ensuring that information is less susceptible to manipulation, organizations can maintain the integrity of online communication channels, reinforce public confidence, and promote informed decision-making. Moreover, the real-time capabilities of these integrated approaches improve societal resilience against coordinated disinformation campaigns, enabling rapid responses to emerging threats.

5. Informing Policy and Governance: Insights derived from integrated analysis inform regulatory frameworks and cybersecurity policies. Governments, platforms, and civil society can use evidence from social graphs and bot network analyses to design effective countermeasures against malicious campaigns, ensuring a safer online environment. Transparency in open-source data collection and analysis also supports accountability and ethical oversight, strengthening the governance of digital information.

6. Strategic Cybersecurity Advantage: From a broader cybersecurity perspective, understanding how influence campaigns propagate and which actors drive them provides a strategic advantage. Organizations and institutions can prioritize protective measures, allocate resources effectively, and anticipate emerging manipulation tactics. This proactive approach reduces vulnerability to both external and internal threats, reinforcing the security of digital infrastructure and information ecosystems.

The rapid proliferation of digital communication platforms has fundamentally changed the landscape of information security. Influence campaigns, coordinated misinformation, and automated amplification through bot networks pose significant risks to the reliability and credibility of information. Open-source data, social graph analysis, and bot network detection collectively address these risks by providing comprehensive visibility into how information spreads, who controls it, and how it can be manipulated.

Open-source data allows organizations and analysts to continuously monitor public digital spaces, detecting abnormal trends, sudden surges in activity, or recurring patterns of misleading content. By aggregating and analyzing this data, it becomes possible to identify potential threats before they escalate into widespread disinformation. This early warning capability is critical for protecting both institutional decision-making and public perception, as rapid intervention can prevent misinformation from gaining traction.

Social graph analysis adds another layer of security by mapping the structural relationships between users and communities. Understanding the network topology helps identify influential nodes, information bottlenecks, and clusters that are being targeted or manipulated. By revealing how narratives propagate and which accounts are central to the spread of information, social graphs enable precise mitigation strategies, such as isolating malicious nodes or disrupting coordinated campaigns, thereby reducing the risk of manipulation at the systemic level.

Bot networks significantly exacerbate threats to information security. Automated accounts can amplify messages, simulate widespread agreement, and obscure the origin of content, making it harder for users and institutions to discern authentic information. Detecting and neutralizing these networks is essential for maintaining the integrity of digital spaces. Integrating bot detection with open-source monitoring and graph-based insights allows for real-time identification of suspicious activity, minimizing the window in which misinformation can influence public discourse.

The combination of these methods enhances trust and resilience across digital ecosystems. Organizations and governments can respond more quickly to threats, users can rely on more credible information, and platforms can maintain safer communication channels. Furthermore, the insights gained from this integrated approach inform cybersecurity policies, regulatory decisions, and public awareness campaigns, reinforcing the overall defense against malicious information operations.

Ultimately, leveraging open-source data, social graphs, and bot network analysis transforms information security from a reactive process into a proactive discipline. By continuously monitoring, mapping, and analyzing digital information flows, stakeholders can anticipate manipulation attempts, mitigate their impact, and safeguard the integrity and credibility of information. This multidimensional approach is becoming increasingly essential as the scale, speed, and sophistication of digital influence operations continue to grow.

In the modern digital environment, information security extends far beyond traditional concerns like data breaches or system intrusions – it increasingly encompasses the integrity, authenticity, and reliability of information itself. Influence campaigns, coordinated misinformation, and manipulative digital content represent a significant and evolving threat to societies, governments, and organizations. The integration of open-source data, social graph analysis, and bot network detection plays a transformative role in addressing these challenges by providing both situational awareness and actionable intelligence.

Open-source data serves as the first line of defense. By continuously monitoring publicly available content across social media platforms, blogs, forums, and other digital channels, analysts can identify early signs of manipulation or coordinated campaigns. Patterns such as repeated messaging, sudden spikes in engagement, or anomalous sentiment shifts act as warning indicators of potential threats. This proactive monitoring is essential for minimizing the impact of disinformation before it spreads widely and influences public perception, corporate decision-making, or political processes.

Social graph analysis enhances security by revealing the underlying structures through which information flows. By mapping nodes (users, accounts) and edges (interactions, shares, mentions), analysts can detect communities susceptible to manipulation, identify central actors amplifying misleading content, and uncover the pathways by which narratives propagate. Understanding these structures allows for targeted interventions: isolating malicious clusters, flagging high-risk

content for review, or disrupting channels used for rapid dissemination of false information. The ability to analyze these networks in real time significantly strengthens the capacity to prevent large-scale information contamination.

Bot networks represent a particularly insidious threat to information security. Automated accounts can generate, amplify, and coordinate content at a scale that far exceeds human capacity. These networks can manufacture the appearance of consensus, manipulate trending topics, and obscure the origins of information, making detection and mitigation increasingly challenging. Integrating bot detection with social graph and open-source data analysis allows for the identification of both individual automated accounts and larger orchestrated networks. Real-time detection is crucial, as it prevents bots from achieving viral reach and limits their potential to influence public discourse or critical decision-making processes.

The convergence of these three analytical approaches – open-source data, social graphs, and bot detection – creates a multidimensional defense framework. It enables stakeholders to not only detect emerging threats but also understand the mechanisms and actors driving them. This holistic perspective enhances resilience against a range of information security challenges, from political misinformation campaigns to corporate espionage, and even public health-related disinformation.

Moreover, these methods inform policy-making and strategic decision-making at national and organizational levels. Governments can use insights from integrated analysis to design regulatory frameworks and countermeasures against coordinated online threats. Corporations can safeguard brand reputation, protect sensitive communications, and maintain customer trust by monitoring digital ecosystems for malicious campaigns. At the societal level, promoting awareness of manipulative practices and empowering users with verified information contributes to a more resilient information environment.

The availability of open-source data has been a transformative factor in the development of large language models (LLMs). These models rely on massive datasets to learn linguistic structures, semantic relationships, and contextual reasoning, enabling them to generate coherent, human-like text and perform complex language tasks. Open-source content – including social media posts, blogs, forums, code repositories, and academic publications – provides the diversity and scale necessary to train models capable of understanding nuanced language patterns across multiple domains.

Open-source data allows LLMs to learn from real-world communication, capturing idiomatic expressions, technical terminology, and evolving language trends. By exposing models to diverse perspectives and writing styles, open-source corpora enhance their generalization capabilities, improving performance across tasks such as question answering, summarization, content generation, and sentiment analysis. Moreover, open-access datasets facilitate reproducibility and collaboration in AI research, allowing independent teams to refine, benchmark, and innovate on LLM architectures without relying solely on proprietary data.

Social graph information further influences LLM development and application. Understanding the relational structure of digital interactions helps models contextualize content within networks, detect patterns of influence, and interpret the propagation of ideas. For example, integrating social graph insights into LLM training or evaluation can improve the detection of coordinated messaging, disinformation campaigns, or emerging narratives, enabling models to provide more informed outputs and assist in monitoring digital ecosystems.

Bot networks also have an indirect yet significant impact on LLMs. Since a substantial portion of online content may be generated or amplified by automated accounts, training LLMs on unfiltered datasets can expose them to biases, repetitive content, or manipulative narratives. Awareness of bot-originated data encourages researchers to implement filtering, weighting, or preprocessing strategies,

ensuring that models learn from high-quality, representative content. Additionally, LLMs can be trained to identify patterns associated with automated or coordinated behavior, making them valuable tools for detecting disinformation and supporting cybersecurity or social media monitoring efforts.

Finally, the interplay between open-source data, social graphs, and bot networks shapes not only the training of LLMs but also their practical applications. Models informed by these datasets can support real-time analysis of information flows, content moderation, influence detection, and threat assessment. This integration enhances the utility of LLMs in domains ranging from digital security to social research, highlighting their growing role as analytical tools for understanding complex, dynamic information environments.

In conclusion, open-source data, social graphs, and bot network insights collectively enhance the development, accuracy, and applicability of LLMs. They provide the scale, diversity, and contextual understanding necessary for models to perform robustly while also informing strategies to mitigate bias, identify coordinated manipulation, and contribute to safer and more transparent digital ecosystems.

Ultimately, the application of open-source intelligence, social graph analysis, and bot network detection transforms information security from a reactive discipline into a proactive one. It allows for the continuous monitoring, evaluation, and mitigation of threats in real time, addressing the rapidly evolving tactics of malicious actors. In an era where digital information flows are central to governance, commerce, and public life, leveraging these technologies is essential for maintaining the integrity, reliability, and trustworthiness of information.

Conclusions. The research presented demonstrates the critical role of open-source data, social graph analysis, and bot network detection in understanding and mitigating influence campaigns within digital ecosystems. Each of these components contributes uniquely to the broader goal of maintaining information integrity, detecting

manipulation, and strengthening digital resilience. Taken together, they provide a comprehensive framework for both academic investigation and practical application in real-time monitoring of online information flows.

1. The Significance of Open-Source Data: Open-source data constitutes the foundational layer for detecting influence campaigns. Publicly accessible content – including social media posts, comments, shares, hashtags, and metadata – offers a rich, continuously updated resource for analysis. By leveraging open-source data, researchers can track emerging trends, identify abnormal behavior, and detect early signs of coordinated activity. The transparency and accessibility of these datasets also facilitate reproducibility and ethical compliance, enabling a wide range of stakeholders, from academic researchers to cybersecurity teams, to develop and deploy detection systems. Importantly, open-source data allows for real-time monitoring, providing early warnings before misinformation or manipulative narratives can significantly influence public opinion or decision-making.

2. The Role of Social Graphs: Social graphs provide a structural perspective, revealing how information travels through networks and identifying nodes and clusters critical to the spread of narratives. Metrics such as centrality, clustering, network density, and modularity allow the detection of influential users, tightly knit communities, and coordinated groups. Temporal and weighted graph analysis enhances this capability by capturing the dynamics of interactions, the strength of relationships, and patterns of synchronized activity indicative of manipulation. Social graph analysis enables not only the identification of coordinated campaigns but also a deeper understanding of how influence propagates, which actors hold disproportionate power, and where interventions may be most effective.

3. The Impact of Bot Networks: Automated accounts constitute a primary mechanism for amplification and manipulation in digital influence campaigns. Bot networks can operate at scale, simulate human behavior, and coordinate across multiple platforms

to generate the appearance of widespread consensus. Detecting these networks is crucial for maintaining the integrity of online discourse. By combining behavioral analysis, content similarity, and network structure evaluation, it is possible to distinguish automated accounts from genuine users, map the scope of coordinated campaigns, and disrupt malicious activity in real time. Machine learning approaches, including graph neural networks and anomaly detection models, provide robust tools for identifying sophisticated bots that adaptively mimic human behavior.

4. **Integration for Real-Time Detection:** The integration of open-source data, social graph analysis, and bot network detection forms a multidimensional and highly effective approach to real-time influence monitoring. Open-source data provides continuous input and contextual content, social graphs reveal structural and relational dynamics, and bot detection identifies automated amplification networks. Together, these components enable a system capable of detecting emerging campaigns at both micro and macro levels, providing actionable intelligence for content moderation, threat mitigation, and strategic decision-making. Real-time integration ensures that interventions occur promptly, reducing the potential impact of misinformation, propaganda, or coordinated manipulation on public discourse.

5. **Implications for Information Security:** From an information security perspective, this integrated approach significantly enhances the resilience, reliability, and trustworthiness of digital ecosystems. Early detection of emerging threats, mapping of coordinated networks, and neutralization of automated amplification mechanisms collectively prevent the widespread dissemination of misleading information. Insights from this approach inform policy-making, platform governance, corporate decision-making, and public awareness initiatives. Furthermore, by monitoring information flows continuously and in real time, organizations and governments can anticipate threats, allocate resources effectively, and mitigate risks to critical infrastructure, public opinion, and societal stability.

6. Strategic and Societal Value: The methodologies examined in this research have far-reaching strategic and societal implications. For governments, they support the protection of democratic processes, election integrity, and national security. For corporations, they enhance brand protection, safeguard customer trust, and prevent reputational damage. For society at large, these tools promote a more informed and resilient public, capable of critically evaluating information and resisting manipulative campaigns. By fostering transparency, accountability, and timely intervention, the integration of open-source intelligence, social graph analysis, and bot detection strengthens trust in digital communication channels and the reliability of information ecosystems.

7. Future Directions: As digital manipulation tactics evolve, continued innovation in data collection, graph analysis, and automated detection is essential. Emerging approaches, such as multi-platform integration, AI-driven content analysis, and predictive modeling of influence campaigns, offer promising avenues for enhancing detection accuracy and response speed. Additionally, ethical considerations, privacy preservation, and bias mitigation must remain central to the development and deployment of these systems, ensuring that the benefits of influence detection do not come at the expense of individual rights or societal trust.

In conclusion, the synergy of open-source data, social graph analysis, and bot network detection represents a transformative framework for understanding and countering influence campaigns. By combining content monitoring, structural network insights, and automated threat identification, this integrated approach enables real-time detection, proactive mitigation, and enhanced information security. Its application not only strengthens digital ecosystems against manipulation but also provides a robust foundation for informed decision-making, policy formulation, and societal resilience in an increasingly interconnected and vulnerable information landscape.

The research highlights that modern information environments are highly interconnected, dynamic, and vulnerable to manipulation. Influence campaigns exploit these characteristics by coordinating messages, leveraging automated accounts, and targeting key communities. Understanding the interplay between open-source data, social graphs, and bot networks is therefore essential for developing effective strategies to maintain the integrity and reliability of digital information.

One of the key insights is that no single method – whether content analysis, network mapping, or bot detection – is sufficient on its own. Only through the integrated use of these approaches can analysts identify both overt and covert operations, trace the flow of influence, and anticipate emerging threats. This multidimensional perspective allows for more accurate detection and more effective mitigation of digital manipulation.

The study also underscores the importance of real-time capabilities. Influence campaigns are most effective when they can spread quickly and exploit viral dynamics. Continuous monitoring and analysis enable timely intervention, preventing false or misleading information from reaching critical mass and minimizing its societal, political, and economic impact.

Finally, the findings emphasize that combating influence campaigns is not solely a technical challenge; it is a multidisciplinary endeavor. Collaboration between data scientists, cybersecurity specialists, social researchers, policymakers, and platform operators is essential to translate analytical insights into actionable strategies. By combining technological innovation with policy frameworks and public awareness, it is possible to build more resilient, secure, and trustworthy information ecosystems.

References

1. Allen, Jennifer, Howland, Baird, Mobius, Markus, Rothschild, David & Watts, Duncan J. Evaluating the fake news problem at the scale of the information ecosystem. Sci. Adv. 6(14).

2. Aseel, Addawood, Adam, Badawy, Kristina, Lerman, and Emilio, Ferrara (2019). Linguistic cues to deception: Identifying political trolls on social media. In Proceedings of the international AAAI conference on web and social media, Vol. 13. 15–25.

3. Bacallao-Pino, L. M. (2016). Radical political communication and social media: The case of the Mexican #yosoy132. In T. Dezelan & I. Vobic (Eds.), (R)evolutionizing political communications through social media (pp. 56–74). Hershey, PA: IGI Global.

4. Bail, C. A. et al. (2017). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late. Proc. Natl. Acad. Sci. 117(1), 243–250.

5. Bessi, A., and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. First Monday 21(11).

6. Enos, Ryan D. & Hersh, Eitan D. (2015). Party activists as campaign advertisers: The ground campaign as a principal – agent problem. Am. Polit. Sci. Rev. 109(2), 252–278.

7. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. (2016). The rise of social bots. Comm. ACM 59(7): 96–104.

8. Guidance AI (2023). Guidance AI Repository. https://github.com/guidance-ai/guidance

9. Isabelle, Augenstein, Timothy, Baldwin, Meeyoung Cha, Tanmoy, Chakraborty, Giovanni, Luca Ciampaglia, David, Corney, Renee, DiResta, Emilio, Ferrara, Scott, Hale, Alon, Halevy, et al. (2023). Factuality Challenges in the Era of Large Language Models. arXiv preprint arXiv:2310.05189.

10. Jason, Wei, Maarten, Bosma, Vincent, Y. Zhao, Kelvin, Guu, Adams Wei Yu, Brian, Lester, Nan, Du, Andrew, M. Dai, and Quoc, V. Le (2021). Finetuned Language Models Are Zero-Shot Learners. CoRR abs/2109.01652. arXiv:2109.01652 https://arxiv.org/abs/2109.01652

11. Julie, Jiang, Xiang, Ren, Emilio, Ferrara, et al. (2021). Social media polarization and echo chambers in the context of COVID-19: Case study. JMIRx med 2, 3, e29570.

12. Kai-Cheng, Yang and Filippo, Menczer (2023). Anatomy of an AI-powered malicious social botnet. arXiv preprint arXiv:2307.16336.

13. Keller, F., Schoch, D., Stier, S. & Yang, J. How to manipulate social media: Analyzing political astroturfing using ground truth data from South

Korea. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, pp. 564–567.

14. Krafft, P. M. & Donovan, Joan. Disinformation by design: The use of evidence collages and platform filtering in a media manipulation campaign. Polit. Commun. 37(2), 194–214.

15. Loomba, Sahil, de Figueiredo, Alexandre, Piatek, Simon J., de Graaf, Kristen & Larson, Heidi J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat. Hum. Behav. 5(3), 337–348.

16. Luca, Luceri, Ashok, Deb, Silvia, Giordano, and Emilio, Ferrara (2019). Evolution of bot and human behavior during elections. First Monday.

17. Luca, Luceri, Silvia, Giordano, and Emilio, Ferrara (2020). Detecting troll behavior via inverse reinforcement learning: A case study of Russian trolls in the 2016 US election. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14. 417–427.

18. Lukito, Josephine et al. The wolves in sheep's clothing: How Russia's internet research agency tweets appeared in U. S. news as vox populi. Int. J. Press/Polit. 25(2), 196–216.

19. Muhammad, Shahid Iqbal Malik, Tahir, Imran, and Jamjoom, Mona, Mamdouh (2023). How to detect propaganda from social media? Exploitation of semantic and fine-tuned language models. PeerJ Computer Science 9, e1248.

20. Stella, M.; Ferrara, E.; and De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. PNAS 115(49): 12435–12440.

21. Yang, K.-C.; Hui, P.-M.; and Menczer, F. (2019). Bot electioneering volume: Visualizing social bot activity during elections. In Companion Proc. WWW Conference, 214–217.

22. Zijian, Cai, Zhaoxuan, Tan, Zhenyu, Lei, Hongrui, Wang, Zifeng, Zhu, Qinghua, Zheng, and Minnan, Luo (2023). LMBot: Distilling Graph Knowledge into Language Model for Graph-less Deployment in Twitter Bot Detection. arXiv preprint arXiv:2306.17408.

CHAPTER 4.
AI IMPLEMENTATION IN UKRAINE: POLICIES, ETHICS, AND ECONOMIC IMPACT

4.1. AI USE: RISKS, ETHICS, PRIVACY, ACCOUNTABILITY

Introduction. The rapid expansion of artificial intelligence (AI) in strategic communications has created an urgent need to examine the ethical and human-rights implications of these technologies. As governments, businesses, and media organizations increasingly rely on AI-driven systems for information dissemination, decision-making, and audience engagement, the potential for both positive impact and significant harm grows. This makes a systematic ethical framework not simply desirable but essential for ensuring that AI serves the public interest rather than undermines it.

One core concern lies in the risks associated with opaque algorithms, data misuse, and automated content generation. AI systems can inadvertently reproduce discrimination, manipulate public opinion, or infringe on privacy through large-scale surveillance. These risks are amplified in strategic communication environments, where messaging can directly influence democratic processes, social stability, and individual autonomy. Understanding the nature of these risks is critical for preserving fundamental human rights such as freedom of expression, equality, and personal dignity.

Equally important is the need for robust harm-minimization strategies. Ethical strategic communication with AI requires transparent data practices, inclusive design, and continual assessment of algorithmic impacts. Building systems that can detect and mitigate

bias, prevent misinformation, and safeguard vulnerable groups is crucial for ensuring that technological innovation does not come at the expense of social justice. Human oversight and multidisciplinary collaboration remain central to responsible AI deployment.

Finally, developing clear protocols and governance mechanisms is vital to institutionalizing ethical practices. This includes establishing regulatory standards, accountability frameworks, and organizational guidelines that define acceptable uses of AI in communication. Such protocols not only help prevent abuse but also build trust among stakeholders, from end-users to policymakers. By embedding human-rights principles into every stage of AI-supported communication, society can better harness the transformative potential of these technologies while upholding the values that underpin democratic life.

Artificial intelligence (AI) is considered a potentially general-purpose technology (GPT) capable of transforming various sectors of the economy and social life. A significant contribution to the formation of this perspective was made by Goldfarb (Goldfarb, 2024) and Bekar et al. (Bekar, C., Carlaw, K., & Lipsey, R., 2018). Goldfarb (Goldfarb, 2024) compares AI to electricity and computers, emphasizing its universality and its ability to stimulate innovation across numerous industries.

Bekar and co-authors (Bekar, C., Carlaw, K., & Lipsey, R., 2018) identify six key characteristics of GPTs, including complementarity with other technologies, the lack of close substitutes, and evolutionary development from simple to complex forms. These features became the theoretical basis for considering AI as a GPT in our study.

Researchers such as Qian et al. (Qian, Y., Siau, K. L., & Nah, F. F., 2024), Horvitz et al. (2019), Goldfarb (Goldfarb, 2024), and others analyze a wide range of consequences of AI adoption: economic, social, ethical, and security-related. They highlight potential risks to the labor market, threats of information manipulation, and challenges associated with autonomous weapon systems. At the same time, Goldfarb (Goldfarb, 2024) and Felten et al. (Felten, E., Raj, M., & Seamans, R., 2024) note the positive potential

of AI, including productivity gains, stimulation of economic growth, and acceleration of innovation.

Regarding public policy, key issues of AI regulation have been studied by Horvitz et al. (Horvitz, E., Conitzer, V., McIlraith, S., & Stone, P., 2024), Ulnicane & Erkkilä (Ulnicane, I., & Erkkilä, T., 2023). Their works emphasize the need for global coordination, the creation of effective legal mechanisms, addressing ethical responsibility issues, and integrating socio-technical narratives into policymaking processes. Particularly valuable is Walter's (Walter, Y., 2024) approach, which proposes the concept of "dynamic laws" – regulatory frameworks capable of adapting to rapid changes in the AI field. Walter (Walter, Y., 2024) also stresses the importance of involving technical experts and ensuring transparency in the development and adoption of new rules.

Large Language Models (LLMs), as a subset of AI, introduce distinct ethical and security challenges within strategic communications. While their ability to generate coherent, contextually relevant, and human-like content offers substantial opportunities for efficiency, personalization, and audience engagement, it simultaneously amplifies the risks of misuse. LLMs can be weaponized to produce persuasive disinformation at scale, subtly manipulate public opinion, and generate content that reinforces existing societal biases. Unlike traditional AI systems, LLMs operate with a high degree of autonomy in natural language generation, making harmful outputs less predictable and harder to monitor.

One critical concern lies in the capacity of LLMs to amplify bias and discrimination embedded in their training data. If unchecked, these models can reproduce stereotypes, marginalize vulnerable groups, and perpetuate systemic inequalities in ways that are difficult to detect. For example, automated generation of news articles, social media posts, or policy briefs can inadvertently reflect skewed perspectives or underrepresented viewpoints, leading to information asymmetries and unequal influence over public discourse.

Another significant threat is large-scale misinformation and disinformation campaigns. LLMs can produce vast volumes of text tailored to specific audiences, potentially generating narratives that mislead, polarize, or manipulate. When used maliciously, these capabilities pose a direct threat to democratic processes, social cohesion, and public trust in institutions. The speed and scale at which LLMs operate outpace traditional fact-checking and moderation mechanisms, making real-time detection and mitigation exceedingly difficult.

Furthermore, the opacity and unpredictability of LLM decision-making complicate accountability. Even well-intentioned deployments may produce harmful or unintended outputs, and tracing responsibility becomes challenging when models operate across multiple communication channels. This raises questions not only about ethical oversight but also about legal liability, particularly in contexts where automated content influences elections, public health messaging, or crisis communication.

To address these threats, it is essential to embed robust harm-minimization and governance mechanisms. Transparent model training, continuous auditing, bias detection, and human-in-the-loop oversight must be integral to LLM deployment in strategic communications. Multidisciplinary collaboration – combining AI expertise, ethics, law, and social sciences – is necessary to anticipate emergent risks and develop effective mitigation strategies. Dynamic monitoring systems capable of detecting manipulative outputs and coordinating rapid intervention are critical for ensuring that the deployment of LLMs aligns with human-rights principles and the public interest.

In summary, while LLMs hold transformative potential for enhancing strategic communication, they simultaneously heighten ethical, societal, and security risks. Failure to address these challenges proactively could result in widespread manipulation of information, erosion of trust in institutions, and infringements on fundamental

human rights such as freedom of expression, equality, and privacy. Therefore, integrating LLM-specific safeguards into AI governance frameworks, public policy, and organizational practices is essential for realizing the benefits of these technologies while minimizing their potential for harm.

The aim of this study is to identify ethical risks associated with the use of AI in strategic communications, including bias, misinformation, and infringement on human rights.

Presentation of the main research material. A theoretical model of AI-related ethical risks is constructed based on digital ethics, human-rights theory, and socio-technical analysis. Concepts such as algorithmic opacity, bias, autonomy, accountability, and privacy serve as core analytical categories. Existing frameworks (AI risk taxonomies, fairness metrics, accountability guidelines) are synthesized to define the primary dimensions of risk: discrimination, misinformation, privacy harm, manipulation, and security threats.

The methodological framework of this study integrates ethical risk assessment with deep learning approaches to analyze AI-driven strategic communication. The research is structured in two main stages: identification and categorization of risks, followed by technical analysis and mitigation strategies using LSTM models.

The first stage focuses on systematically identifying and classifying ethical risks associated with AI in strategic communication. Drawing on digital ethics, human-rights theory, and socio-technical analysis, risks are grouped into the following categories:

1. Algorithmic bias and discrimination – risks of reproducing or amplifying social inequalities through automated decision-making.
2. Misinformation and manipulative content – dissemination of false or emotionally manipulative messages affecting public opinion.
3. Privacy violations – unauthorized collection, analysis, or leakage of personal data.
4. Lack of transparency and accountability – opaque algorithms that hinder human oversight.

5. Security threats – risks associated with autonomous systems and malicious content propagation.

Data sources for this stage include publicly available datasets of news content, social media posts, corporate communication records, and platform transparency reports. Each instance is annotated manually or semi-automatically according to risk type, severity, and affected stakeholders. This provides a structured corpus for subsequent technical analysis.

The second stage employs Long Short-Term Memory (LSTM) neural networks to analyze the textual content of strategic communication and detect potential ethical risks. The approach includes the following steps:

1. Data Preprocessing:

– Text tokenization, lemmatization, and removal of stop words.

– Conversion of text into numerical representations using embeddings (e.g., Word2Vec, GloVe, or contextual embeddings).

2. LSTM Model Construction:

– A sequence-based LSTM architecture is designed to capture the temporal and contextual dependencies of language in communication messages.

– The model is trained to classify messages according to risk type (bias, misinformation, manipulative intent, privacy violation).

3. Risk Scoring and Interpretation:

– Model outputs are mapped to risk scores representing the likelihood and severity of ethical concern.

– Attention mechanisms or feature importance methods are applied to interpret which textual elements contribute most to identified risks.

4. Validation and Evaluation:

– Model performance is evaluated using standard metrics (accuracy, F1-score, precision, recall) on annotated datasets.

– Cross-validation ensures generalization and robustness across diverse communication scenarios.

5. Integration into Ethical Governance Framework:

– The results from LSTM analysis inform risk-mitigation strategies, including content moderation, bias reduction, and privacy safeguards.

– Human oversight and policy guidelines are applied in tandem to ensure responsible deployment of AI in communication contexts.

Research Results. There are numerous studies dedicated to identifying the ethical risks of AI use, among which the works of Douglas (Douglas, D. M., Lacey, J. & Howard, D. Ethical, 2024), Matthew G. Hanna (Matthew G. Hanna, Liron Pantanowitz, Brian Jackson, 2025), Stockman, C. (Stockman, 2024) etc., should be highlighted.

Table 4.1 – Ethical and Legal Risks of AI in Strategic Communications

Risk Name	Brief Description	Mitigation Methods
1	2	3
Opacity (Opaque Algorithms)	Lack of clarity in model functioning, making it difficult to detect errors, manipulation, and discrimination; reduces accountability of developers	Explainable AI (XAI), transparent algorithm documentation, model audits, human-in-the-loop
Discrimination and Algorithmic Bias	AI may reproduce or amplify social inequalities; risk of violating equality principles	Regular bias testing, training data correction, fairness metrics, inclusive design
Public Opinion Manipulation	Automated content generation can influence political processes, shape public sentiment, and increase polarization	Fact-checking, generative content control, monitoring message dissemination, ethical AI frameworks
Mass Privacy Invasion	Use of AI for monitoring, big data analysis, or covert surveillance; threatens privacy rights	Differential privacy, data minimization, encryption, strict data handling policies
Spread of Disinformation and Fakes	Generative models can produce plausible but false content; accelerates dissemination of harmful content	NLP-based fake news detection, fact-checking, media literacy promotion, source verification

End of Table 4.1

1	2	3
Lack of Effective Control and Accountability	Insufficient regulations and international coordination; difficulty assigning responsibility for AI-induced harm	Development of regulatory standards, protocols, ethical codes, institutionalized auditing and reporting
Undermining Autonomy and Freedom of Expression	Manipulative algorithms may limit information access or influence individual choices	Transparent recommendation algorithms, source diversity, human-in-the-loop, protection of informed decision-making rights
Security Threats	Use of AI in autonomous systems (weapons, targeted attacks, content tampering)	Cybersecurity measures, anomaly monitoring, safe deployment protocols, limiting autonomy of critical systems

Source: compiled by the authors

The analysis of ethical and legal risks associated with AI in strategic communications highlights the multifaceted challenges posed by emerging technologies. Key concerns include algorithmic opacity, bias, manipulation of public opinion, privacy violations, and security threats. These risks demonstrate that AI systems, while offering significant potential for efficiency and innovation, can also inadvertently undermine fundamental human rights, social equality, and democratic processes if not properly governed. The diversity of risks emphasizes the need for a comprehensive approach that considers both technical vulnerabilities and societal impacts.

Effective mitigation requires a combination of technological, organizational, and regulatory measures. Transparency-enhancing tools such as Explainable AI, rigorous bias testing, and human-in-the-loop oversight are critical for reducing algorithmic harm. Complementary strategies include fact-checking, monitoring of content dissemination, data minimization, cybersecurity protocols, and the development of clear accountability frameworks. By integrating these measures

into AI deployment, organizations can balance innovation with ethical responsibility, safeguarding individual rights and promoting trust in AI-mediated communication.

As a countermeasure, AI methods can be used to reduce risks. Among the datasets for training ML and NN models, it is worth noting the LIAR Dataset, Enron Email, FakeCovid Fact-Checked News Dataset, etc., which are available on the Kaggle platform (Kaggle). For training AI detection, it is advisable to use the MiRAGeNews datasets, Community Forensics, etc.

As a result of analyzing the approaches presented on Kaggle, the use of SVM, Logistic Regression, Random Forest, LSTM, etc., can be highlighted. For example, a simple LSTM model:

```
from tensorflow.keras.layers import Input, Embedding, Dropout, Bidirectional, LSTM, Dense
from tensorflow.keras.models import Model
inputs = Input(shape=(max_length,))
x = Embedding(input_dim=vocab_size,
output_dim=embedding_dim,
weights=[embedding_matrix],
trainable=False)(inputs)
x = Dropout(0.2)(x)
x = Bidirectional(LSTM(n, return_sequences=True, dropout=0.2, recurrent_dropout=0.2))(x)
x = Bidirectional(LSTM(n, dropout=0.2, recurrent_dropout=0.2))(x)
x = Dense(n, activation='relu')(x)
x = Dropout(0.5)(x)
outputs = Dense(1, activation='sigmoid')(x)
)
```

with 4 layers achieved an accuracy of 0.705 on the validation data.

However, particularly high results can be achieved by using an ensemble of models together with standard ML models and pre-trained NN modules, reaching accuracy of up to 0.9. For example:

```
stack_train = np.vstack((nn_pred_train, lr_pred_train))
stack_test = np.vstack((nn_pred_test, lr_pred_test))

stack_inputs = Input(shape=(2,))
x = Dense(4, activation='relu')(stack_inputs)
x = Dense(2, activation='relu')(x)
stack_outputs = Dense(1, activation='sigmoid')(x)
```

To address the ethical risks associated with AI in strategic communications, a combination of technological, organizational, and regulatory measures is essential. One key approach involves leveraging AI itself to detect and mitigate potential harms. Techniques such as Natural Language Processing (NLP) and Long Short-Term Memory (LSTM) neural networks can identify biased, manipulative, or false content in real time, allowing organizations to proactively intervene before harm spreads. Using ensemble models that combine multiple machine learning and deep learning approaches enhances detection accuracy, providing a robust toolset for ethical oversight.

Human oversight remains a critical component of effective countermeasures. Incorporating human-in-the-loop mechanisms ensures that AI-driven decisions are reviewed by trained professionals, preventing automated systems from reinforcing biases or inadvertently spreading misinformation. Additionally, transparency-enhancing tools such as Explainable AI (XAI) allow stakeholders to understand model decisions, fostering accountability and trust. Regular audits, bias testing, and monitoring of content dissemination complement these technical solutions by ensuring continuous alignment with ethical standards.

Organizational and regulatory strategies further reinforce these measures. Establishing clear protocols, ethical guidelines, and accountability frameworks helps define acceptable uses of AI in communication, while regulatory coordination at national and international levels ensures consistent enforcement. Data protection practices, including encryption, data minimization, and differential

privacy, safeguard individual rights and mitigate privacy risks. Together, these countermeasures create a comprehensive ethical governance framework that balances innovation with social responsibility, ensuring that AI supports democratic processes and human rights rather than undermining them.

Conclusions. The analysis of AI in strategic communications underscores both the transformative potential of these technologies and the significant ethical risks they pose. Key concerns include algorithmic opacity, bias, misinformation, privacy violations, and threats to security, which can collectively undermine human rights, democratic processes, and social trust. While AI offers opportunities for efficiency, innovation, and improved audience engagement, its misuse or unregulated deployment can exacerbate social inequalities and manipulate public opinion, highlighting the urgent need for ethical oversight.

Effective mitigation of these risks requires a multifaceted approach that combines technological, organizational, and regulatory strategies. Techniques such as Explainable AI, bias testing, human-in-the-loop systems, content monitoring, and secure data practices are critical to reducing harm. Additionally, the use of AI itself, through machine learning and neural networks like LSTM models, can aid in detecting misinformation, manipulative content, and bias, demonstrating that AI can be both a source of risk and a tool for risk management. Ensemble modeling and integration with traditional machine learning approaches further enhance the accuracy and reliability of ethical risk detection.

Ultimately, the responsible deployment of AI in strategic communications depends on embedding ethical principles, human-rights considerations, and governance frameworks throughout the development and implementation process. Regulatory standards, accountability mechanisms, and transparent organizational policies are essential for ensuring that AI supports social good rather than undermines it. By combining technical solutions with human oversight and multidisciplinary collaboration, organizations can harness AI's

benefits while safeguarding individual rights, promoting trust, and sustaining democratic and social integrity.

References

1. Goldfarb, A. (2024). Pause artificial intelligence research? Understanding AI policy challenges. Canadian Journal of Economics, 57(2), 363–377. https://doi.org/10.1111/caje.12705

2. Bekar, C., Carlaw, K., & Lipsey, R. (2018). General purpose technologies in theory, application and controversy: a review. Journal of Evolutionary Economics, 28(5). https://doi.org/10.1007/s00191-017-0546-0

3. Qian, Y., Siau, K. L., & Nah, F. F. (2024). Societal impacts of artificial intelligence: Ethical, legal, and governance issues. Societal Impacts, 3, 100040. https://doi.org/10.1016/J.SOCIMP.2024.100040a

4. Horvitz, E., Conitzer, V., McIlraith, S., & Stone, P. (2024). Now, Later, and Lasting: 10 Priorities for AI Research, Policy, and Practice. Communications of the ACM, 67(6), 39–40. https://doi.org/10.1145/3637866

5. Felten, E., Raj, M., & Seamans, R. (2024). Generative AI Requires Broad Labor Policy Considerations. Communications of the ACM, 67(8), 29–32. https://doi.org/10.1145/3637864/ASSET/5FB620B0-5F2E-4FFE-992A6844360C94D4/ASSETS/GRAPHIC/3637864_FIG01H.JPG

6. Ulnicane, I., & Erkkilä, T. (2023). Politics and policy of Artificial Intelligence. Review of Policy Research, 40(5), 612–625. https://doi.org/10.1111/ropr.12574

7. Walter, Y. (2024). Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation – A contemporary overview and an analysis of socioeconomic consequences. Discover Artificial Intelligence 2024 4:1, 4(1), 1–24. https://doi.org/10.1007/S44163-024-00109-4

8. Douglas, D. M., Lacey, J. & Howard, D. Ethical risk for AI. *AI Ethics* 5, 2189–2203 (2025). https://doi.org/10.1007/s43681-024-00549-9

9. Matthew, G. Hanna, Liron, Pantanowitz, Brian, Jackson, Octavia, Palmer, Shyam, Visweswaran, Joshua Pantanowitz, Mustafa Deebajah, Hooman H. Rashidi, Ethical and Bias Considerations in Artificial Intelligence/Machine Learning, Modern Pathology, Volume 38, Issue 3, 2025, 100686, ISSN 0893-3952, https://doi.org/10.1016/j.modpat.2024.100686

10. Stockman, C. (2024). Generative AI and the Ethical Risks Associated with Human-Computer Symbiosis. *Weizenbaum Journal of the Digital Society*, *5*(1). https://doi.org/10.34669/wi.wjds/5.1.2
11. Platform Kaggle. https://Kaggle.com

4.2. FROM DEFENSE TO DEVELOPMENT: A ROADMAP FOR AN AI ECOSYSTEM FOR INFORMATION SECURITY AND ECONOMIC GROWTH (2026–2030)

Introduction. In the coming decade, artificial intelligence (AI) will become a defining force shaping both national security and economic competitiveness. As digital infrastructures expand and cyber threats grow in scale and sophistication, information security can no longer rely solely on reactive or defensive measures. Instead, it must evolve into a proactive, innovation-driven ecosystem where security technologies also serve as engines of economic development. The transition from defense-oriented frameworks to development-oriented strategies marks a critical turning point for governments, industries, and research institutions worldwide.

The period from 2026 to 2030 presents a unique opportunity to build a comprehensive AI ecosystem that simultaneously strengthens information security and stimulates sustainable economic growth. Advances in machine learning, data analytics, and automation enable not only more resilient cybersecurity architectures but also the creation of new markets, high-value jobs, and competitive digital industries. By integrating AI into security governance, infrastructure, and human capital development, nations can transform cyber defense capabilities into scalable platforms for innovation.

This roadmap outlines a strategic shift from isolated security solutions toward an interconnected AI-driven ecosystem. It emphasizes the alignment of policy, technology, education, and investment to ensure that AI enhances trust, resilience, and productivity across the digital economy. Moving from defense to development is not merely a technological evolution – it is a strategic imperative for long-term security, economic growth, and digital sovereignty in the 2026–2030 horizon.

Beyond technological advancement, the successful formation of an AI-driven ecosystem for information security requires a coordinated institutional and regulatory foundation. Fragmented initiatives, isolated pilot projects, and short-term defensive investments are insufficient to address systemic cyber risks or unlock AI's full economic potential. A long-term roadmap must therefore prioritize cross-sector collaboration between the public sector, private industry, academia, and international partners. Such collaboration ensures that AI solutions are interoperable, ethically grounded, and aligned with broader economic and security objectives.

Equally important is the development of human capital and organizational capacity. AI-based information security systems depend not only on algorithms and data, but also on skilled professionals capable of designing, governing, and continuously improving these systems. Investment in education, reskilling, and applied research is essential to create a workforce that can operate at the intersection of cybersecurity, data science, and economic innovation. From specialized training programs to research-driven innovation hubs, human capital development serves as a cornerstone of a resilient AI ecosystem.

The roadmap for 2026–2030 also recognizes that trust is a critical enabling factor. The widespread adoption of AI in information security will depend on transparent governance models, robust data protection mechanisms, and clear accountability frameworks. Without public and institutional trust, even the most advanced AI technologies risk limited

adoption or societal resistance. Therefore, ethical AI principles, risk management standards, and regulatory harmonization must evolve in parallel with technological deployment.

Finally, the transition from defense to development reframes information security as a strategic economic asset rather than a cost center. AI-powered security solutions can drive productivity across industries, enable secure digital transformation, and support the growth of data-driven services and platforms. By embedding security into the foundations of digital growth, this roadmap envisions an ecosystem where resilience, innovation, and economic competitiveness reinforce one another. In this context, information security becomes not only a shield against threats, but a catalyst for sustainable development and long-term prosperity.

Presentation of the main research material. Large Language Models (LLMs) represent a foundational technology within the next generation of AI-driven ecosystems, with transformative implications for both information security and economic development. Their ability to process, generate, and reason over vast volumes of unstructured data positions them as critical enablers of intelligence-driven security and scalable digital innovation. Within the 2026–2030 horizon, LLMs are expected to evolve from experimental tools into core infrastructure components of national and sectoral AI ecosystems.

In the domain of information security, LLMs significantly enhance situational awareness and decision-making capabilities. By analyzing threat intelligence reports, security logs, vulnerability disclosures, and real-time incident data, LLM-powered systems can accelerate threat detection, automate incident response, and support security analysts with contextualized insights. Unlike traditional rule-based systems, LLMs enable adaptive security operations by understanding intent, correlating disparate signals, and generating actionable recommendations. This shift reduces response times, lowers operational costs, and improves resilience against complex and evolving cyber threats.

LLMs also play a critical role in strengthening secure software development and digital infrastructure. Integrated into development pipelines, they can assist in code review, vulnerability identification, secure-by-design architecture, and compliance verification. This capability not only improves the baseline security of digital products and services but also enhances productivity across the software and IT sectors. As a result, security becomes embedded within innovation processes rather than applied as a reactive control, supporting both risk reduction and economic efficiency.

From an economic perspective, LLMs act as multipliers of value creation across industries. They enable the development of new AI-powered security services, knowledge-based platforms, and data-driven business models. Small and medium-sized enterprises, in particular, benefit from LLM-based tools that lower barriers to entry by providing access to advanced analytics, automated compliance, and intelligent decision support. This democratization of AI capabilities fosters entrepreneurship, stimulates innovation ecosystems, and contributes to inclusive economic growth.

At the ecosystem level, LLMs facilitate interoperability and knowledge sharing across institutions. By serving as interfaces between humans and complex digital systems, they enable more effective collaboration among policymakers, researchers, security professionals, and businesses. LLM-powered knowledge systems can support policy modeling, regulatory impact analysis, and strategic forecasting, aligning security investments with long-term economic objectives.

However, the integration of LLMs also introduces new security, ethical, and governance challenges. Risks such as model misuse, data leakage, adversarial manipulation, and dependency on opaque systems must be proactively addressed. Therefore, the roadmap emphasizes the development of sovereign and trusted LLM infrastructures, robust evaluation and auditing mechanisms, and clear accountability frameworks. Responsible deployment of LLMs is essential to ensure that their economic and security benefits outweigh associated risks.

In summary, LLMs are not merely tools within the AI ecosystem – they are strategic assets that reshape how information security and economic growth reinforce one another. When governed effectively, LLMs enable a transition from fragmented defensive measures to integrated, intelligence-driven ecosystems that support resilience, innovation, and sustainable development.

Large Language Models (LLMs) are emerging as one of the most influential technological layers within modern AI ecosystems, fundamentally reshaping the relationship between information security, digital governance, and economic growth. Unlike narrow AI systems designed for isolated tasks, LLMs operate as general-purpose cognitive infrastructures capable of understanding context, synthesizing knowledge, and supporting complex decision-making across domains. Their integration into the AI ecosystem between 2026 and 2030 will mark a qualitative shift from tool-based automation toward intelligence-driven systems that unify security, productivity, and innovation.

1. LLMs as Cognitive Infrastructure for Information Security.

In the field of information security, LLMs function as cognitive amplifiers that enhance the ability of organizations to perceive, interpret, and respond to cyber risks. Modern security environments generate massive volumes of heterogeneous data, including logs, alerts, threat intelligence feeds, policy documents, and regulatory requirements. LLMs enable semantic integration of these data sources, transforming fragmented information into coherent situational awareness.

LLM-powered security platforms can:

- Interpret complex attack patterns and adversarial behaviors.
- Correlate technical signals with geopolitical, economic, and organizational contexts.
- Generate human-readable explanations for security decisions and risks.
- Support analysts in hypothesis generation and strategic threat modeling.

This capability significantly reduces cognitive overload for security professionals and allows a transition from reactive incident handling to anticipatory and predictive security operations. As a result, information security evolves from a purely technical discipline into a strategic intelligence function embedded in organizational governance.

2. Automation and Augmentation of Security Operations.

LLMs play a critical role in automating and augmenting Security Operations Centers (SOCs). Integrated into SOC workflows, LLMs can triage alerts, summarize incidents, propose mitigation strategies, and generate post-incident reports. Importantly, this automation is not limited to technical actions but extends to decision support and knowledge transfer.

By reducing reliance on scarce expert labor, LLMs help address the global cybersecurity skills gap. They enable junior analysts to operate at higher levels of effectiveness and allow senior experts to focus on complex, high-impact problems. This human – AI collaboration model improves operational resilience while lowering costs, creating positive spillover effects for economic productivity.

3. Secure-by-Design Digital Economy Enabled by LLMs.

Beyond operational security, LLMs fundamentally reshape how secure digital systems are designed and built. When embedded in software development lifecycles, LLMs support:

- Secure coding practices and automated vulnerability detection.
- Compliance with security standards and regulations by design.
- Continuous security validation throughout system evolution.

This leads to a structural reduction of systemic cyber risk across the digital economy. Secure-by-design approaches lower the cost of breaches, increase trust in digital services, and enable faster innovation cycles. As trust becomes a competitive advantage, economies that successfully integrate LLM-driven security into digital infrastructure gain stronger positions in global value chains.

4. Economic Growth Through LLM-Driven Security Innovation.

LLMs act as catalysts for new markets and business models at the intersection of AI, security, and knowledge services. They enable the emergence of:

- AI-powered managed security services.
- Automated compliance and governance platforms.
- Industry-specific security intelligence solutions.
- Exportable AI security products and services.

For small and medium-sized enterprises, LLMs dramatically lower entry barriers by providing access to advanced security and analytics capabilities without requiring large in-house teams. This democratization of security technology stimulates entrepreneurship, supports innovation clusters, and accelerates the diffusion of digital technologies across the economy.

At the macroeconomic level, reduced cyber risk and increased trust in digital systems translate into higher investment attractiveness, more efficient digital trade, and stronger integration into global digital markets.

5. LLMs as Interfaces Between Policy, Technology, and Markets.

One of the most transformative roles of LLMs lies in their function as interfaces between complex systems. LLMs can translate policy objectives into technical requirements, map regulatory constraints onto operational processes, and help decision-makers evaluate trade-offs between security, innovation, and economic growth.

In public governance, LLM-based systems support.

- Strategic forecasting and scenario analysis.
- Policy impact assessment for AI and cybersecurity regulations.
- Coordination across ministries, agencies, and critical infrastructure operators.

This capability enables more adaptive, evidence-based policymaking and reduces the lag between technological change and regulatory response. As a result, governance itself becomes more resilient and innovation-friendly.

6. Sovereignty, Trust, and Strategic Autonomy.

Despite their benefits, LLMs introduce new strategic dependencies and risks. Control over training data, model architectures, and deployment infrastructures becomes a matter of national and economic security. Dependence on opaque or externally controlled LLMs can expose ecosystems to data leakage, manipulation, or supply-chain risks.

Therefore, a key element of the roadmap is the development of trusted and, where necessary, sovereign LLM capabilities. This includes:

- Transparent model governance and auditability.
- Robust safeguards against misuse and adversarial attacks.
- Alignment with ethical AI principles and human oversight.

Trustworthy LLMs are essential not only for security but also for sustained economic adoption. Without trust, the scaling potential of LLM-driven ecosystems remains limited.

7. Long-Term Systemic Impact (2026–2030).

Between 2026 and 2030, LLMs will increasingly function as systemic enablers rather than isolated technologies. Their cumulative impact will manifest in:

- Higher resilience of digital infrastructure.
- Reduced economic losses from cyber incidents.
- Increased productivity and innovation capacity.
- Stronger alignment between security investments and economic outcomes.

In this context, LLMs transform information security from a defensive cost into a strategic driver of growth. They enable an ecosystem where intelligence, trust, and innovation reinforce one another, supporting sustainable development in an increasingly digital and interconnected world.

While Large Language Models (LLMs) offer substantial benefits for strengthening information security and enabling economic growth, their large-scale deployment introduces a new class of systemic, technical, economic, and governance risks. These risks

are not peripheral; they directly affect trust, resilience, and long-term sustainability of the AI ecosystem. Without proactive risk management, LLMs may amplify existing vulnerabilities, create new attack surfaces, and undermine both security objectives and economic outcomes.

1. Security and Adversarial Risks.

LLMs expand the cyber threat landscape by introducing novel attack vectors. Adversarial actors may exploit LLMs through prompt injection, data poisoning, model extraction, or indirect prompt manipulation via compromised data sources. In security-critical environments, such attacks can lead to misinformation, incorrect decision-making, or unauthorized disclosure of sensitive information.

Moreover, LLMs can be weaponized to automate phishing, social engineering, malware generation, and reconnaissance activities at scale. This dual-use nature creates an asymmetry: the same capabilities that enhance defensive intelligence can also empower attackers, increasing the speed and sophistication of cyber threats.

If not properly secured, LLM-integrated systems risk becoming high-value targets themselves, potentially cascading failures across interconnected digital infrastructures.

2. Data Privacy and Confidentiality Risks.

LLMs rely heavily on large volumes of data, often including sensitive, proprietary, or regulated information. Improper handling of training data, prompts, or outputs can result in unintended data leakage, violation of privacy regulations, and loss of intellectual property.

In enterprise and government settings, the use of externally hosted or opaque LLM services increases exposure to cross-border data transfer risks and compliance uncertainties. Even when models are deployed locally, inference-time data retention and logging practices can introduce latent vulnerabilities.

Failure to address data governance risks undermines public trust and may lead to regulatory backlash, limiting economic scalability of AI-driven solutions.

3. Reliability, Explainability, and Decision Risk.

LLMs are probabilistic systems that may generate inaccurate, misleading, or fabricated outputs (hallucinations). In the context of information security, such errors can have severe consequences, including misclassification of threats, inappropriate response actions, or flawed policy recommendations.

The limited explainability of LLM decision processes complicates accountability and auditability. When LLMs are embedded in automated or semi-automated security workflows, errors may propagate rapidly across systems before human oversight can intervene.

This creates a risk of overreliance on AI-generated intelligence, particularly in high-pressure or resource-constrained environments.

4. Systemic Dependency and Concentration Risk.

The AI ecosystem faces growing dependency on a small number of large-scale LLM providers and proprietary model architectures. Such concentration creates single points of failure and reduces strategic autonomy for organizations and states.

Vendor lock-in, limited transparency, and restricted control over model updates may expose critical infrastructure and economic systems to external shocks, including geopolitical tensions, supply-chain disruptions, or unilateral changes in service conditions.

From an economic perspective, excessive concentration can distort markets, suppress innovation, and limit the development of local AI ecosystems.

5. Ethical, Social, and Workforce Risks.

LLMs may encode biases present in training data, leading to discriminatory or unfair outcomes in security assessments, access control, or risk profiling. In information security, biased models may disproportionately flag or overlook certain actors, behaviors, or regions, undermining both effectiveness and legitimacy.

At the workforce level, rapid automation enabled by LLMs may displace certain roles while increasing demand for high-skilled positions. Without proactive reskilling and transition

strategies, this imbalance may exacerbate inequality and create resistance to AI adoption.

Ethical misalignment between LLM behavior and societal values poses long-term risks to social trust and institutional legitimacy.

6. Governance and Accountability Gaps.

The integration of LLMs into complex decision-making systems challenges existing governance frameworks. Ambiguity around responsibility – whether for model developers, deployers, or users – complicates incident response, liability, and legal enforcement.

Current regulatory regimes often lag behind technological capabilities, creating gray zones where risks are insufficiently addressed. Overregulation, however, may stifle innovation and reduce economic competitiveness.

Striking the right balance between control and flexibility remains a central governance challenge for the AI ecosystem.

7. Long-Term Systemic and Strategic Risks.

Over time, poorly governed LLM deployment may lead to erosion of human expertise, reduced critical thinking, and excessive automation of strategic judgment. In security-sensitive domains, this could weaken institutional resilience and adaptive capacity.

At a strategic level, misalignment between LLM development trajectories and national economic or security priorities may result in path dependency, limiting future policy options.

The risks associated with LLMs are systemic rather than isolated and must be addressed at the ecosystem level. Effective risk management requires coordinated action across technology design, governance, regulation, education, and international cooperation. Only by embedding safeguards, transparency, and human oversight into the AI ecosystem can LLMs serve as trusted enablers of information security and sustainable economic growth (see Table 4.2? p. 172–173).

The risk analysis demonstrates that Large Language Models represent not only a powerful enabler of information security and economic growth, but also a source of high-impact systemic risks.

Table 4.2 – Risk Matrix: Large Language Models in Information Security and Economic Growth

Risk Category	Risk Description	Impact on Information Security	Impact on Economic Growth	Strategic Relevance (2026–2030)
1	2	3	4	5
Adversarial Attacks	Prompt injection, data poisoning, model extraction, adversarial manipulation	Compromised threat detection, false intelligence, leakage of sensitive data	Increased costs of incidents, reduced trust in AI-driven services	High
Weaponization of LLMs	Use of LLMs for phishing, social engineering, malware generation	Escalation of attack scale and sophistication	Higher cybercrime losses, pressure on digital markets	High
Data Privacy & Confidentiality	Leakage of personal, classified, or proprietary data through training or inference	Regulatory violations, loss of sensitive information	Legal risks, reduced investment attractiveness	High
Reliability & Hallucinations	Inaccurate or fabricated outputs used in decision-making	Incorrect incident response, misclassification of threats	Inefficient resource allocation, operational failures	Medium – High
Lack of Explainability	Limited transparency of model reasoning	Reduced auditability, accountability gaps	Barriers to adoption in regulated industries	Medium
Overreliance on Automation	Excessive trust in AI outputs without human oversight	Degradation of human expertise, systemic failures	Long-term productivity and resilience risks	Medium – High
Systemic Dependency	Dependence on a small number of proprietary LLM providers	Loss of strategic autonomy, supply-chain vulnerabilities	Market concentration, reduced innovation	High

End of Table 4.2

1	2	3	4	5
Vendor Lock-in	Limited portability and interoperability of LLM solutions	Reduced flexibility in security architecture	Increased costs, reduced competitiveness	Medium
Bias and Discrimination	Embedded biases in training data and outputs	Unfair or ineffective security profiling	Social resistance, reputational damage	Medium
Workforce Disruption	Automation-driven displacement without reskilling	Skills gaps in critical security roles	Inequality, slower adoption of AI solutions	Medium
Governance & Liability Gaps	Unclear responsibility for AI-driven decisions	Ineffective incident response and enforcement	Legal uncertainty, chilling effect on innovation	High
Geopolitical & Sovereignty Risks	Cross-border control of models, data, and infrastructure	Exposure to external pressure or manipulation	Strategic vulnerability of digital economy	High
Long-Term Strategic Drift	Misalignment between LLM deployment and national priorities	Erosion of institutional resilience	Path dependency limiting future growth options	Medium – High

Source: compiled by the authors

The most critical risks identified in the matrix are characterized by their cross-cutting nature: they simultaneously affect security operations, economic stability, governance structures, and long-term strategic autonomy. This confirms that LLM-related risks cannot be addressed through isolated technical controls and require an ecosystem-level response.

A key conclusion is that security-related risks dominate the high-impact category. Adversarial attacks, model weaponization, and data confidentiality breaches are assessed as high-risk due to their potential to scale rapidly and propagate across interconnected systems. These risks directly undermine trust in AI-driven security solutions, which is a foundational prerequisite for their economic adoption. Without effective safeguards, the same LLM capabilities that enhance defensive intelligence may accelerate offensive cyber activities, creating a net negative security outcome.

The table also highlights systemic dependency and concentration risks as strategic threats rather than operational issues. Dependence on a limited number of proprietary LLM providers introduces vulnerabilities related to vendor lock-in, loss of strategic autonomy, and exposure to geopolitical dynamics. From an economic perspective, this concentration risks distorting markets and constraining the development of domestic AI ecosystems, thereby limiting long-term innovation and competitiveness.

Another important finding is the interdependence between reliability, explainability, and governance risks. Hallucinations, lack of transparency, and unclear accountability structures collectively increase the likelihood of erroneous or untraceable decisions in security-critical contexts. These factors elevate operational risk and create regulatory uncertainty, particularly in highly regulated sectors, which may slow investment and adoption despite clear productivity gains.

The analysis further indicates that human and organizational risks remain significant. Overreliance on automation and insufficient workforce adaptation threaten the sustainability of AI-enabled security

systems. While LLMs can mitigate skills shortages in the short term, failure to invest in human capital development may erode institutional expertise and long-term resilience, weakening both security outcomes and economic performance.

Finally, the table underscores that long-term strategic risks are cumulative rather than immediate. Strategic drift, erosion of sovereignty, and misalignment with national priorities may not produce immediate failures but can significantly constrain future policy options and economic trajectories. These risks emphasize the importance of aligning LLM deployment with broader security, industrial, and innovation strategies over the 2026–2030 period.

Building an AI Ecosystem for Information Security and Economic Growth (2026–2030).

1. Strategic Vision.

The strategic objective of this roadmap is to transform information security from a reactive, defense-oriented function into a proactive driver of innovation, trust, and sustainable economic growth. Between 2026 and 2030, artificial intelligence – particularly Large Language Models – will serve as a foundational layer of a resilient AI ecosystem that integrates security, economic development, governance, and human capital.

This strategy envisions information security not as a constraint on digital growth, but as an enabling infrastructure that increases productivity, accelerates innovation, and strengthens digital sovereignty. The transition "from defense to development" reflects a paradigm shift: security investments are leveraged to create long-term economic value, competitive advantage, and institutional resilience.

2. Strategic Principles.

The roadmap is guided by the following core principles:

1) Security by Design, Not by Reaction AI-driven security must be embedded into digital systems, platforms, and services from inception rather than applied post hoc;

2) Ecosystem-Level Thinking Effective outcomes require coordination across technology, policy, markets, education, and international cooperation;

3) Trust as Economic Infrastructure Trustworthy AI, data governance, and transparency are prerequisites for large-scale adoption and investment;

4) Human-Centered AutomationAI augments human decision-making rather than replacing accountability and strategic judgment;

5) Strategic Autonomy and Openness The ecosystem balances openness and interoperability with sovereignty, resilience, and risk control.

3. Strategic Pillars.

Pillar I: AI-Driven Information Security Infrastructure.

Objective: Build resilient, intelligent, and interoperable security infrastructure powered by AI and LLMs.

Key actions:

- Deploy AI-enabled threat detection, prediction, and response platforms across critical sectors.
- Integrate LLMs into Security Operations Centers (SOCs) for intelligence synthesis and decision support.
- Establish shared threat intelligence and secure data exchange mechanisms.
- Promote secure-by-design digital architectures and AI-assisted secure software development.

Expected outcomes:

- Reduced cyber incident impact and response time.
- Lower systemic cyber risk across the digital economy.
- Scalable security capabilities accessible to organizations of all sizes.

Pillar II: Trusted and Sovereign AI Foundations.

Objective: Ensure trust, control, and accountability in the deployment of LLMs and AI systems.

Key actions:

- Develop trusted LLM deployment models (including sovereign or hybrid architectures).

- Implement AI auditing, evaluation, and lifecycle governance frameworks.
- Enforce robust data governance, privacy, and confidentiality standards.
- Align AI development with ethical principles and human oversight.

Expected outcomes:

- Increased trust in AI-driven security and governance systems.
- Reduced dependency on opaque or external AI infrastructures.
- Regulatory certainty that supports innovation and investment.

Pillar III: Economic Development and Innovation Enablement.

Objective: Leverage AI-powered security as a catalyst for economic growth and competitiveness.

Key actions:

- Support AI security startups, innovation hubs, and public – private partnerships.
- Enable AI-powered managed security services and compliance platforms.
- Lower adoption barriers for SMEs through shared AI security infrastructure.
- Promote export-oriented AI security solutions and standards.

Expected outcomes:

- Creation of new markets and high-value jobs.
- Increased productivity and digital transformation across industries.
- Strengthened position in global digital value chains.

Pillar IV: Human Capital and Institutional Capacity.

Objective: Build a workforce and institutions capable of governing and operating AI-driven ecosystems.

Key actions:

- Invest in education and reskilling at the intersection of AI, cybersecurity, and policy.
- Develop interdisciplinary research and training centers.
- Promote AI literacy among decision-makers and regulators.

- Institutionalize human – AI collaboration models.

Expected outcomes:

- Reduced skills gaps in critical security domains.
- Stronger institutional resilience and adaptive capacity.
- Sustainable long-term use of AI technologies.

Pillar V: Governance, Policy, and International Cooperation.

Objective: Create adaptive governance frameworks that align security, innovation, and economic goals.

Key actions:

- Establish cross-sector AI and cybersecurity coordination bodies.
- Integrate AI risk management into national economic and security planning.
- Align standards and regulations with international frameworks.
- Promote international cooperation on AI security norms and resilience.

Expected outcomes:

- Faster policy adaptation to technological change.
- Reduced regulatory fragmentation and uncertainty.
- Enhanced cross-border trust and digital trade.

4. Phased Implementation Roadmap (2026–2030).

Phase 1 (2026–2027): Foundation.

- Establish governance frameworks and risk controls.
- Pilot AI and LLM-based security platforms.
- Launch workforce development initiatives.

Phase 2 (2028–2029): Scaling.

- Expand AI security infrastructure across sectors.
- Integrate LLMs into core digital and governance processes.
- Support ecosystem-wide adoption by SMEs and public institutions.

Phase 3 (2030): Consolidation.

- Optimize ecosystem performance and resilience.
- Align AI security capabilities with long-term economic strategy.
- Institutionalize continuous improvement and innovation cycles.

5. Strategic Impact.

By 2030, the successful implementation of this strategy will result in:

- A resilient AI ecosystem where information security enables growth.
- Reduced economic losses from cyber risks.
- Increased trust in digital systems and AI governance.
- Stronger alignment between technological innovation and national development goals.

6. Strategic Conclusion.

The transition from defense to development represents a structural transformation in how societies approach information security and economic growth. Artificial intelligence – anchored by Large Language Models – serves as both the technological and strategic catalyst of this transformation. When governed responsibly and deployed strategically, AI enables a future where security is not a constraint, but a foundation for sustainable innovation, competitiveness, and prosperity in the digital age.

Pillar I: AI-Driven Information Security Infrastructure.

Strategic Goal: Enhance cyber resilience and operational efficiency through AI-enabled security systems.

Table 4.3 – Pillar I: AI-Driven Information Security Infrastructure

KPI	Metric Definition	Measurement Method	Target by 2030
AI-enabled threat coverage	Share of critical systems protected by AI-based security tools	% of systems	≥85%
Mean Time to Detect (MTTD)	Average time to identify incidents	Minutes / hours	–60% vs 2025
Mean Time to Respond (MTTR)	Average time to contain incidents	Minutes / hours	–50% vs 2025
False positive rate	Incorrect alerts generated by security systems	% of alerts	≤10%
Cross-sector threat intelligence sharing	Number of organizations connected to shared AI platforms	Absolute / % growth	≥3× growth

Source: compiled by the authors

Pillar II: Trusted and Sovereign AI Foundations.

Strategic Goal: Ensure trust, transparency, and strategic autonomy in AI and LLM deployment.

Table 4.4 – Pillar II: Trusted and Sovereign AI Foundations

KPI	Metric Definition	Measurement Method	Target by 2030
Trusted LLM adoption	Share of AI systems using audited / certified LLMs	% of deployments	≥70%
AI audit coverage	Systems subject to regular AI risk and ethics audits	% of systems	≥90%
Data governance compliance	Compliance with data protection and sovereignty standards	Audit score	≥95%
Dependency concentration index	Share of AI usage tied to top 3 providers	%	≤50%
Incident traceability	AI-related incidents with clear accountability	%	≥95%

Source: compiled by the authorscompiled by the authors

Pillar III: Economic Development and Innovation Enablement.

Strategic Goal: Use AI-driven security to stimulate innovation, productivity, and market growth.

Table 4.5 – Pillar III:
Economic Development and Innovation Enablement

KPI	Metric Definition	Measurement Method	Target by 2030
AI security market growth	Annual growth rate of AI security sector	% CAGR	≥15%
SME AI adoption rate	SMEs using AI-based security solutions	% of SMEs	≥60%
AI security startups	Number of active AI security startups	Absolute	≥2× vs 2025
Export share	Share of AI security exports in total digital exports	%	≥25%
Productivity uplift	Efficiency gains in AI-enabled organizations	% increase	≥20%

Source: compiled by the authors

Pillar IV: Human Capital and Institutional Capacity.

Strategic Goal: Develop skilled workforce and institutions capable of governing AI ecosystems.

Table 4.6 – Pillar IV: Human Capital and Institutional Capacity

KPI	Metric Definition	Measurement Method	Target by 2030
AI – security specialists	Number of certified professionals	Absolute	≥2.5× vs 2025
Reskilling participation	Workforce involved in AI/ cyber upskilling	% of workforce	≥40%
AI literacy (leadership)	Decision-makers trained in AI governance	%	≥80%
Human – AI collaboration index	Tasks involving AI with human oversight	%	≥90%
Skills gap index	Unfilled critical AI-security positions	%	≤10%

Source: compiled by the authors

Pillar V: Governance, Policy, and International Cooperation.

Strategic Goal: Align AI security governance with economic strategy and international norms.

Table 4.7 – Pillar V: Governance, Policy, and International Cooperation

KPI	Metric Definition	Measurement Method	Target by 2030
Policy alignment index	Coherence between AI, security, and economic policies	Composite score	≥85/100
Regulatory response time	Time to update regulations after tech shifts	Months	≤12
International standards adoption	Alignment with global AI/cyber frameworks	% compliance	≥90%
Cross-border cooperation	Active international AI security agreements	Absolute	≥2× vs 2025
Public trust in AI	Public confidence in AI governance	Survey-based index	≥75%

Source: compiled by the authors

Table 4.8 – Cross-Cutting Meta-KPIs (Ecosystem-Level)

Meta-KPI	Description	Target by 2030
Cyber loss reduction	Reduction in economic losses from cyber incidents	–40%
AI ROI	Economic return on AI security investments	≥3:1
Ecosystem resilience index	Ability to absorb and recover from shocks	≥85/100
Trust & adoption correlation	Correlation between trust and AI adoption	Positive & increasing

Source: compiled by the authors

Implementation Notes:

• KPIs should be reviewed annually and adjusted based on technological and geopolitical changes.

• Metrics combine quantitative, qualitative, and composite indicators.

• Independent audits and public reporting increase transparency and trust.

The implementation of an AI-driven ecosystem for information security and economic growth in Ukraine faces profound and multidimensional challenges due to the ongoing military aggression, disrupted infrastructure, and heightened cybersecurity threats. While the strategic roadmap envisions a transition "from defense to development," the operational environment in Ukraine imposes severe constraints that require adaptation, prioritization, and resilience-oriented planning.

Ukraine operates under a constant threat environment, with state-sponsored cyberattacks, ransomware campaigns, and disinformation operations targeting critical infrastructure, government systems, and private enterprises. These persistent attacks complicate the deployment of AI-driven security systems, as resources must be primarily directed toward immediate defense needs. As a result, initiatives focused on long-term development, such as establishing AI research hubs,

innovation platforms, and advanced LLM-based security systems, face delays and limitations in scale.

The physical and digital infrastructure in Ukraine is also under significant strain. Military operations and targeted strikes have damaged energy grids, communication networks, and data centers, creating unstable environments for AI deployment. Reliable connectivity, computational resources, and secure cloud platforms–essential for the effective operation of AI and LLM systems–are often compromised. Consequently, strategies must emphasize mobile, decentralized, and resilient architectures capable of functioning even under disrupted conditions.

Human capital represents another critical bottleneck. Skilled IT and cybersecurity professionals are frequently displaced due to conflict, leading to a reduction in the workforce required for AI system development, deployment, and governance. Continuous professional development programs and university-level education initiatives face operational interruptions, further constraining workforce readiness. To address this, human capital strategies must leverage remote learning, diaspora engagement, and rapid upskilling programs adaptable to crisis conditions.

Economic and financial constraints exacerbate these challenges. A substantial portion of public budgets is directed toward military expenditure and emergency response, limiting investments in AI infrastructure, research, and innovation. Simultaneously, businesses confront market disruptions, reduced investment capacity, and economic uncertainty, which diminishes private-sector participation in the AI ecosystem. In this context, international financial support, grants, and public – private partnerships are critical to sustain development initiatives in the near term.

Governance and policy adaptation are similarly affected by the exigencies of conflict. Security imperatives dominate political decision-making, creating tension between long-term AI strategy and immediate defense priorities. Developing adaptive AI governance

frameworks becomes particularly challenging when institutional capacity is focused on survival rather than innovation. Agile, modular governance models are therefore essential, allowing AI deployment to scale rapidly while maintaining oversight and regulatory compliance.

Trust and adoption present additional challenges. Persistent information warfare, disinformation campaigns, and cyberattacks erode public confidence in digital systems and AI solutions. Citizens and businesses may be hesitant to adopt AI-driven services if they perceive increased vulnerability to cyberattacks or potential misuse. Building trust, therefore, requires transparent deployment, robust security, clear communication, and user-focused engagement to ensure both societal acceptance and practical adoption of AI technologies.

Despite these obstacles, opportunities remain for implementing a resilient and impactful AI ecosystem. Prioritizing the deployment of decentralized, mobile-capable, and offline-resilient AI and LLM systems can mitigate the effects of disrupted infrastructure. International partnerships provide access to funding, expertise, and technology transfer, enabling Ukraine to overcome financial and technical constraints. Human capital initiatives that focus on remote learning and diaspora engagement can address workforce gaps while maintaining continuity in AI development. Critically, integrating short-term defensive measures with phased introduction of AI-driven economic initiatives allows the country to balance immediate security needs with long-term growth objectives. Simultaneously, emphasizing transparency, auditable AI systems, and proactive communication can strengthen trust, counter disinformation, and facilitate the adoption of AI solutions across sectors.

In conclusion, the implementation of an AI ecosystem roadmap in Ukraine under conditions of Russian aggression is exceptionally challenging due to the combination of high-intensity cyber threats, disrupted infrastructure, limited human capital, and constrained financial resources. Nonetheless, by adopting resilience-focused strategies, leveraging international support, and carefully balancing

immediate security imperatives with long-term development goals, Ukraine can gradually build an AI ecosystem capable of enhancing both national security and sustainable economic growth even amid ongoing conflict.

Implementing an AI-driven ecosystem for information security and economic growth in Ukraine under the conditions of ongoing Russian aggression represents one of the most complex challenges for any modern nation-state. The very nature of an AI ecosystem – reliant on stable infrastructure, reliable data flows, skilled human capital, and institutional governance – is fundamentally at odds with the uncertainty, disruption, and accelerated risk environment imposed by active military conflict. Nevertheless, Ukraine's efforts to move from a defensive posture toward a development-oriented AI ecosystem are critical not only for national security but also for economic resilience and long-term technological sovereignty.

The first major challenge arises from the scale and sophistication of cyber threats. Ukrainian networks, critical infrastructure, and digital institutions have been repeatedly targeted by state-sponsored cyberattacks. These attacks range from highly coordinated intrusions into government and financial networks to widespread disinformation campaigns aimed at eroding public trust. In such an environment, the deployment of AI-driven security platforms and Large Language Models (LLMs) must contend with active adversarial interference, making operational stability a core concern. The need for continuous defensive operations often monopolizes cybersecurity resources, delaying initiatives that would otherwise focus on proactive innovation, research, or economic growth. Consequently, the roadmap must prioritize solutions that can operate under persistent threat conditions while gradually introducing development-oriented capabilities.

Compounding this challenge is the disruption of Ukraine's digital and physical infrastructure. Targeted strikes on energy grids, communication networks, and data centers have created unstable and fragmented digital environments. AI systems, particularly those reliant

on cloud computing, high-volume data processing, and continuous connectivity, face difficulties in such a context. Any deployment must therefore incorporate redundancy, decentralization, and offline capabilities to maintain effectiveness during network interruptions or localized infrastructure failures. This requires rethinking conventional AI architectures and adapting them for high-resilience, crisis-compatible deployment scenarios.

Human capital limitations further constrain implementation. The conflict has caused displacement of skilled IT professionals, cybersecurity experts, and data scientists, resulting in gaps in the workforce essential for AI system development and governance. Educational institutions and research centers face operational interruptions, slowing the cultivation of new talent. Addressing this requires innovative approaches, including leveraging the Ukrainian diaspora for remote expertise, implementing accelerated upskilling programs, and developing modular education initiatives that can operate even in disrupted conditions. Human – AI collaboration must be designed to compensate for skill gaps while maintaining high standards of oversight and decision-making quality.

Financial and economic pressures present additional obstacles. Substantial portions of national budgets are allocated to military operations and emergency response, leaving limited resources for AI infrastructure, research, or innovation programs. Businesses face market volatility, reduced access to capital, and operational disruptions, further constraining private-sector participation in AI ecosystem development. In this environment, international support – from European Union programs, NATO initiatives, and multilateral financial mechanisms – becomes essential. Public – private partnerships and targeted grant programs are critical to sustaining development-oriented projects while balancing immediate defense priorities.

Governance and policy challenges in Ukraine under conflict conditions are also significant. Decision-making processes are

dominated by urgent security imperatives, often at the expense of long-term strategic planning. Developing adaptive regulatory frameworks for AI and cybersecurity is difficult when institutional capacity is stretched thin by immediate operational needs. Yet governance frameworks are crucial: without clear accountability, standards, and oversight mechanisms, AI deployment risks becoming fragmented, opaque, or ineffective. Agile, modular policy models that allow rapid iteration, oversight, and integration with international norms are therefore necessary to maintain both security and developmental objectives.

Public trust and social acceptance form an additional layer of complexity. Widespread information warfare and disinformation campaigns erode confidence in digital systems and AI solutions, particularly in the context of ongoing attacks. Citizens and organizations may resist adopting AI-enabled platforms if they perceive increased exposure to cyber threats or misuse. Building trust requires transparent, auditable, and human-centered AI systems, coupled with active public communication strategies that highlight security, ethical governance, and tangible benefits of adoption.

Despite these formidable challenges, Ukraine possesses unique opportunities to implement a resilient AI ecosystem even amid conflict. Prioritizing high-resilience architectures that can function in decentralized and intermittent environments allows AI systems to remain operational under duress. Strategic international partnerships provide access to technology, funding, and expertise, reducing dependency on domestic resources while fostering alignment with global standards. Human capital strategies, including remote engagement, diaspora collaboration, and modular training programs, can mitigate the effects of displacement and skill gaps. Importantly, by integrating immediate defensive measures with phased, development-oriented initiatives, Ukraine can gradually shift toward an AI ecosystem that strengthens economic growth, technological innovation, and national resilience.

In this context, the transition from defense to development is not merely a technological challenge but a profound strategic imperative. It requires balancing immediate security needs with long-term investments in infrastructure, human capital, and governance. By leveraging AI and LLMs as both protective tools and enablers of economic growth, Ukraine can build an ecosystem that withstands ongoing threats while laying the foundation for sustainable innovation, digital sovereignty, and competitive advantage. In essence, success depends on the ability to transform adversity into an opportunity, turning the pressures of conflict into a catalyst for accelerated modernization and strategic resilience (see Table 4.9, p. 189).

The implementation of an AI ecosystem in Ukraine under conditions of Russian aggression requires a phased, adaptive approach that balances immediate security imperatives with long-term economic development. The first phase, Foundation & Resilience (2026–2027), focuses on establishing secure, decentralized, and resilient AI and LLM infrastructure capable of operating under disrupted conditions. This includes the deployment of AI-enabled Security Operations Centers (SOCs) for critical sectors such as government, energy, and finance, alongside the initiation of workforce upskilling programs. Governance and audit frameworks are introduced early to ensure accountability, risk monitoring, and ethical compliance. KPIs in this phase emphasize coverage and operational efficiency, such as achieving AI threat coverage of at least 50%, reducing mean time to respond to incidents by 30%, and training 30% of the cybersecurity workforce. Key risks include active cyber attacks, infrastructure instability, and resource constraints, which are mitigated through resilient system architectures, redundant infrastructures, and strategic prioritization of critical assets.

The second phase, Scaling & Integration (2028–2029), expands AI deployment across public and private sectors while incorporating development-oriented initiatives. LLMs are integrated into predictive threat intelligence and decision-support systems, allowing

Table 4.9 – Phased Implementation Plan: AI Ecosystem for Information Security and Economic Growth in Ukraine (2026–2030)

Phase	Timeline	Strategic Focus	Key Actions / Initiatives	KPI & Metrics	Risks & Mitigation
Phase 1: Foundation & Resilience	2026–2027	Establishing secure and resilient AI infrastructure	Deploy decentralized AI and LLM systems in critical sectors; Pilot AI-enabled SOCs; Initiate workforce upskilling programs; Set initial governance and audit frameworks	AI threat coverage ≥50%; MTTR reduction 30%; 30% workforce trained	High-intensity cyber attacks, infrastructure instability; Mitigation: resilient, offline-capable systems, redundant architectures, priority resource allocation
Phase 2: Scaling & Integration	2028–2029	Expand AI deployment across sectors and integrate development focus	Scale AI security platforms to SMEs and public institutions; Deploy LLMs in predictive intelligence; Launch innovation hubs and AI startups support programs; Strengthen data governance	AI adoption by SMEs ≥40%; Threat intelligence sharing growth 2x; 70% systems audited	Human capital gaps, regulatory lag; Mitigation: remote training, modular policy frameworks, diaspora engagement
Phase 3: Consolidation & Optimization	2030	Align AI ecosystem with national strategy for security and economic growth	Optimize AI system performance; Institutionalize AI governance; Integrate security with economic development; Measure ecosystem-wide impact	AI ROI ≥3:1; Cyber loss reduction ≥40%; Ecosystem resilience index ≥85/100; Public trust ≥75%	Strategic dependency, social resistance; Mitigation: sovereign AI deployment, transparent communication, international partnerships

Source: compiled by the authors

organizations to transition from reactive security to proactive, intelligence-driven operations. Innovation hubs and support programs for AI startups are launched to stimulate entrepreneurship, technological innovation, and economic growth. Metrics focus on ecosystem adoption, with targets such as 40% AI adoption among SMEs, doubling the volume of cross-sector threat intelligence sharing, and auditing 70% of deployed systems. Risks in this phase include workforce shortages, regulatory lag, and limited private-sector engagement. Mitigation strategies include remote training programs, modular and agile policy frameworks, and engagement with the Ukrainian diaspora to supplement human capital needs.

The final phase, Consolidation & Optimization (2030), aims to align the AI ecosystem with long-term national strategy for both security and economic growth. AI systems are optimized for performance and interoperability, governance structures are institutionalized, and security investments are explicitly linked to economic outcomes. Key performance indicators include achieving an AI ROI of at least 3:1, reducing economic losses from cyber incidents by 40%, attaining an ecosystem resilience index of 85/100, and securing public trust levels above 75%. Strategic risks, such as dependency on external providers and societal resistance to AI adoption, are mitigated through sovereign AI deployments, transparent communication, and international partnerships to enhance technological autonomy.

Throughout all phases, the phased implementation plan emphasizes resilience, adaptability, and integration. Each stage builds upon the previous, ensuring that immediate defensive needs do not preclude long-term developmental goals. By aligning AI and LLM deployment with governance frameworks, workforce development, infrastructure resilience, and economic stimulation, Ukraine can gradually transition from a security-focused posture to a growth-oriented AI ecosystem capable of withstanding ongoing threats while creating competitive advantages in the digital economy.

Large Language Models (LLMs), as a powerful subset of AI, present a range of unique risks in the context of strategic communications. Their ability to generate coherent, contextually nuanced, and human-like text at scale offers significant opportunities for efficiency and engagement, but it also amplifies the potential for misuse. Unlike traditional AI tools, LLMs operate with a high degree of autonomy in language generation, making their outputs unpredictable and sometimes difficult to control. This unpredictability introduces ethical and security challenges that can have profound implications for public trust, social stability, and democratic processes.

One of the primary concerns is the amplification of bias and discrimination embedded in training data. LLMs can unintentionally reproduce societal stereotypes, marginalize vulnerable populations, or favor certain perspectives over others. In strategic communications, this may manifest in the automated generation of news articles, social media content, or policy briefs that reflect skewed viewpoints, creating information asymmetries and unequal influence over public discourse. Such biased outputs not only undermine fairness but also threaten the integrity of information ecosystems on which societies rely.

Equally critical is the risk of large-scale misinformation and disinformation. LLMs can produce massive volumes of persuasive content tailored to specific audiences, facilitating manipulation of public opinion and the spread of false narratives. In democratic contexts, these capabilities can be weaponized to influence elections, polarize communities, or distort policy debates. The sheer speed and scale of LLM-generated content can overwhelm traditional fact-checking mechanisms, allowing harmful narratives to propagate before they can be countered effectively.

The opacity of LLM decision-making further exacerbates these risks. Outputs are generated through complex probabilistic patterns that are often inscrutable even to experts, complicating accountability. Even when deployed with good intentions, LLMs can produce unintended or harmful messages, raising difficult questions about

legal and ethical responsibility. Determining who is accountable – the developer, the deploying organization, or the operator – becomes particularly challenging when LLMs operate across multiple communication channels or in automated workflows.

To mitigate these threats, it is essential to implement robust harm-minimization strategies. Transparent model training, continuous auditing, bias detection, and human-in-the-loop oversight are crucial to ensuring responsible deployment. Multidisciplinary collaboration involving AI experts, ethicists, legal scholars, and social scientists is necessary to anticipate emerging risks and to design effective safeguards. Systems must be capable of detecting manipulative outputs, countering misinformation, and protecting vulnerable groups, while embedding human-rights principles into every stage of AI-supported communication.

In conclusion, while LLMs have the potential to transform strategic communications by enhancing productivity, personalization, and engagement, they simultaneously elevate ethical, societal, and security risks. Without proactive governance and careful oversight, these models could facilitate widespread manipulation of information, erode public trust, and infringe on fundamental human rights such as freedom of expression, equality, and privacy. Ensuring that LLMs serve the public interest rather than undermine it requires a comprehensive framework of ethical standards, technical safeguards, and regulatory mechanisms that prioritize transparency, accountability, and social justice.

Conclusions. The analysis of an AI-driven ecosystem for information security and economic growth in Ukraine demonstrates that while the challenges are significant, the strategic opportunities are equally substantial. The ongoing military aggression and persistent cyber threats create an environment of acute operational pressure, infrastructure instability, and human capital limitations. Nevertheless, a carefully phased, resilience-oriented strategy can enable Ukraine to transition from a purely defensive posture toward

a development-focused AI ecosystem that simultaneously strengthens national security, digital sovereignty, and economic competitiveness.

A key conclusion is that resilience must underpin all phases of implementation. The immediate security environment dictates that AI systems, particularly Large Language Models (LLMs), be deployed in decentralized, offline-capable, and fault-tolerant architectures. Without this foundational resilience, long-term development initiatives risk failure under persistent disruption. The phased plan demonstrates that building operational robustness first provides the necessary platform for subsequent scaling, integration, and economic optimization.

Another critical insight is the interdependence of human capital, governance, and technological deployment. The success of AI and LLM systems depends not only on the availability of hardware and software but also on the presence of skilled professionals and institutions capable of managing complex, high-stakes systems. Ukraine's ongoing conflict has created workforce gaps and disrupted educational pipelines, highlighting the importance of remote training, diaspora engagement, and modular upskilling programs. Governance frameworks, in turn, must be agile, transparent, and aligned with international norms to ensure trust, compliance, and accountability. Without coherent human and institutional capacity, the most advanced AI systems would fail to deliver intended security and economic outcomes.

The analysis also underscores the dual-use nature of AI and LLM technologies, which can amplify both defensive capabilities and offensive threats. In Ukraine's context, adversarial actors are highly active, and the risk of model exploitation, misinformation, and systemic manipulation is significant. Therefore, risk mitigation – through secure architectures, ethical AI frameworks, continuous monitoring, and transparency – is not optional but integral to strategy execution. KPIs and metrics outlined in the roadmap provide measurable ways to track resilience, adoption, operational efficiency, and trust, ensuring that progress toward development objectives is continuously monitored and adjusted.

Economic considerations are central to the roadmap's long-term impact. AI-driven security is positioned not as a cost center but as a catalyst for innovation, market creation, and productivity gains. Even under conflict, strategic investments in AI startups, innovation hubs, and SME adoption programs can generate high-value jobs, stimulate entrepreneurship, and strengthen Ukraine's position in global digital markets. Importantly, aligning security infrastructure with economic objectives transforms resilience into competitive advantage, demonstrating that robust defense and economic growth are not mutually exclusive but mutually reinforcing.

Finally, the phased implementation plan highlights the necessity of balancing short-term defense with long-term development. Phase 1 focuses on establishing operational resilience and immediate threat mitigation. Phase 2 emphasizes scaling and integration, linking security capabilities to economic innovation. Phase 3 consolidates achievements, optimizes performance, and institutionalizes governance while reinforcing alignment with national development priorities. This phased approach ensures that immediate threats do not preclude strategic growth, while providing a structured pathway to gradually build a mature AI ecosystem capable of withstanding crisis conditions.

In conclusion, the experience of Ukraine illustrates that building an AI ecosystem in conflict conditions requires a holistic, multi-dimensional approach. Success depends on integrating resilient technology, skilled human capital, adaptive governance, risk management, and economic innovation into a coherent strategy. When executed effectively, this approach allows Ukraine not only to defend its digital sovereignty but also to leverage AI as a tool for economic resilience and long-term growth. Ultimately, the roadmap demonstrates that even in the most challenging operational environments, a carefully designed and monitored AI ecosystem can transform defense imperatives into a foundation for sustainable development, competitive advantage, and national resilience.

References

1. Alom, Z., Carminati, B., Ferrari, E. Detecting Spam Accounts on Twitter. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018.

2. Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., Menczer, F. The spread of low-credibility content by social bots. Nat. Commun. 2018, 9, 1–9.

3. Posetti, J. News industry transformation: Digital technology, social platforms and the spread of misinformation and disinformation. In Journalism,'Fake News' and Disinformation: A Handbook for Journalism Education and Training; UNESCO: Paris, France, 2018.

4. Teyssou, D., Leung, J. M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., Papadopoulou, O., Mezaris, V. The InVID plug-in: Web video verification on the browser. In Proceedings of the First International Workshop on Multimedia Verification, Mountain View, CA, USA, 27 October 2017; pp. 23–30.

5. Marinova, Z.; Spangenberg, J.; Teyssou, D.; Papadopoulos, S.; Sarris, N.; Alaphilippe, A.; Bontcheva, K. Weverify: Wider and enhanced verification for you project overview and tools. In Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–4.

Izdevniecība “Baltija Publishing”
Avotu iela 8 k-1 - 25, Rīga, LV-1011
E-mail: office@baltijapublishing.lv

Iespiests tipogrāfijā SIA “Izdevniecība “Baltija Publishing”
Parakstīts iespiešanai: 2025. gada 23. decembris
Tirāža 300 eks.