PHYSICAL AND MATHEMATICAL SCIENCES

GENERATIVE EDUCATIONAL CONTENT: AUTOMATIC VERIFICATION AND FACT-CHECKING

Sergiy Yevseyev[1]

DOI: https://doi.org/10.30525/978-9934-26-568-6-4

Introduction. «Artificial Intelligence» today is one of the most important aspects of the modern technological world. Introduced to the general public a couple of years ago, Large Language Models (LLMs) and their respective commercial-based implementations, such as ChatGPT, Claude AI, Google Gemini and many, many others – opened a door to various implementations, sometimes changing the landscape dramatically [1]. Teachers are considering how to treat AI-powered responses from students. Is this a violation of academic purity? Or is it just a time-saver in case the student is already familiar with the subject?

Another aspect is how teachers can incorporate AI-powered applications into their existing workflows [2]. It's a well-known fact that LLM can produce incorrect results (even if they sound dangerously convincing). Verify by hand, an old-fashioned way? It was already pointed out that for educational purposes, results of ChatGPT and other LLM-powered tools should be validated – without exception – by some sort of external validation: either human or automatic [3].

This paper proposes that the solution to verifying AI-generated content paradoxically lies in AI itself – specifically, in a novel "panel of experts" verification system utilizing multiple independent LLMs. Through strategic deployment of multiple models with different architectures and training datasets, we can significantly reduce the probability of undetected factual errors and provide more robust content evaluation than single-model approaches currently permit.

Automatic content evaluation challenges. Ultimately, "grading" or evaluating the AI-generated (or student-submitted, AI-assisted) content highlights potentially inaccurate, misleading, or inadequately supported sections. Developing such an automatic verification tool presents a formidable yet essential multi-faceted challenge. It requires a sophisticated interplay of various AI disciplines and a deep understanding of pedagogical needs.

[1] Ivano-Frankivsk National Technical University of Oil and Gas, Ukraine
ORCID: https://orcid.org/0009-0007-9821-8997

One possible solution is creating a "panel of experts" or a "second opinion" system using a set of *independent, highly-trained* LLMs.

Panel of experts. While all LLMs can hallucinate, the specific content and nature of their hallucinations often differ due to variations in their training data, architecture, and fine-tuning. It's less likely that three independent models will *all* converge on the exact same specific factual error, especially if they have different knowledge cut-off dates or primary training focuses.

Each LLM processes information differently and internally represents knowledge. One might catch a nuance or have access to information that another missed or misinterpreted.

If multiple LLMs independently corroborate a piece of information or an assessment, confidence in that information increases. Conversely, disagreement is a strong signal for caution.

Implementation architecture. The proposed verification system operates through a three-layer architecture:

− *Layer 1*: *content decomposition.* The input content (whether AI-generated or student-submitted) undergoes automated decomposition into discrete, verifiable units.

− *Layer 2: multi-model verification*. Each decomposed unit is independently evaluated by at least two different, diverse LLMs, with different architectures (e.g., transformer-based models of varying sizes) and training datasets.

− *Layer 3: consensus analysis and reporting*. The system aggregates model responses through weighted voting based on model confidence and reliability.

Providing quality feedback. Developers should remember that the "verifier LLMs" are not infallible**:** this is the fundamental issue. We are using imperfect tools to check an imperfect tool. The goal is to *reduce* error, not eliminate it entirely using this method alone.

Specialized prompts should require verifier LLMs to explain their reasoning and, if possible, provide a confidence score for their assessment, for e.g.: "Please rate the accuracy of this statement on a scale of 1-5 and explain your rating."

If the content is highly specialized (e.g., advanced quantum physics), general-purpose LLMs might struggle. Future iterations might involve routing verification tasks to LLMs fine-tuned on specific domains, if available and reliable.

If a verifier LLM gives an ambiguous answer, the system could be designed to ask follow-up clarifying questions, thus implementing *iterative querying*.

For truly critical factual claims, the LLM panel's consensus should ideally still be cross-referenced with external, reliable knowledge bases (databases, curated encyclopedias, scientific papers). The LLM panel can help *identify what to check* and *provide an initial assessment*.

Decomposition for verification. As LLMs as a tool are not ideal, it is necessary to *adjust* complex responses for automatic verification. The original content needs to be broken down into smaller, verifiable units (e.g., individual claims, arguments, evidence-conclusion links). *Agreement thresholds* should be introduced: if 2 out of 3 (or N-1 out of N) verifier LLMs agree on a specific point (e.g., "Claim X is false"), that carries significant weight.

Analyzing disagreements. Disagreements are just as important, if not more so. *Why* do they disagree? Does one LLM have more recent information? Is there genuine ambiguity in the statement? Strong disagreements, or cases where all LLMs express low confidence, should be automatically flagged for human review. Over time, if one verifier LLM consistently proves more reliable for certain types of verification tasks, its "vote" or assessment could be weighted more heavily. This requires ongoing performance monitoring.

Meta-reasoning and independent observation. When verifier LLMs provide explanations, the system (or another LLM acting as a "meta-reviewer") could compare these justifications. Are they coherent? Do they cite similar reasons or evidence? If two LLMs provide similar, well-reasoned justifications for an assessment, and a third differs wildly with a weak or nonsensical justification, it might be possible to discount the third.

Conclusion. Using multiple LLMs for verification offers a potentially interesting pathway for reducing manual workload. It doesn't solve the "ground truth" problem entirely, as LLMs can still collectively be wrong or miss nuances, especially on very new or highly specialized information. However, it significantly *reduces the probability of unflagged errors* and provides a more robust system than relying on a single AI model. The key will be the sophistication of the prompting strategies for the verifier LLMs and the intelligence of the adjudication mechanism that synthesizes their outputs [4]. This approach turns the problem into one of managing and interpreting a "committee" of AIs, which is a step forward in robust AI-assisted evaluation.

References:

1. Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models." arXiv preprint. https://arxiv.org/abs/2102.02503

2. Mollick, E. R., & Mollick, L. (2023). "Assigning AI: Seven Approaches for Students, with Prompts." SSRN Electronic Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4475995

3. Kasneci, E., Seßler, K., Küchemann, S. et al. (2023). "ChatGPT for good? On opportunities and challenges of large language models for education." Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

4. Liu, Y., Yin, D., et al. (2023). "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment." arXiv preprint. https://arxiv.org/abs/2303.16634