A model for improving the accuracy of educational content created by generative AI

Oleh V. Talaver[1], Tetiana A. Vakaliuk[1,2,3,4]

[1]*Zhytomyr Polytechnic State University, 103 Chudnivsyka Str., Zhytomyr, 10005, Ukraine*
[2]*Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine*
[3]*Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*
[4]*Academy of Cognitive and Natural Sciences, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*

Abstract

Advancements in artificial intelligence (AI) are reshaping education, enabling personalized and adaptive learning experiences, yet ensuring the reliability of AI-generated content remains a critical challenge. This study addresses this gap by developing a system for text processing and factual claims verification, focusing on extracting factual claims, retrieving evidence from authoritative sources, verifying content, and rewriting it to ensure accuracy while maintaining pedagogical effectiveness. The system is designed to complement manual peer review processes by providing detailed annotations and evidence-based notes related to facts retrieved from different sources. The proposed system employs a multi-layered approach to content verification. At its core, it focuses on three key processes: (1) extracting and refining factual claims from text using sophisticated natural language processing techniques, (2) retrieving and analyzing evidence from multiple authoritative sources, and (3) verifying and rewriting content that will be proposed as a hint to ensure accuracy while maintaining pedagogical effectiveness. It integrates prompt engineering, multi-stage evidence analysis, different source information retrieval, and vector-based embeddings to get and classify evidence as supporting, contradicting, or neutral, using weighted credibility scoring and a proposed decision certainty level for each verification outcome. By employing a multi-layered approach, the system offers a practical and scalable solution for enhancing AI-generated content verification, paving the way for reliable applications in education and organizational knowledge management, and enabling educators and content creators to make informed decisions about content revision.

Keywords

generative AI, education content creation, prompt engineering, factual validation, model hallucinations, verify-and-edit frameworks

1. Introduction

Artificial intelligence (AI) advancements are reshaping every facet of life, with education standing at the forefront of this transformation [1, 2]. By integrating subject matter experts (SMEs) directly into the process, AI can potentially enhance the accessibility, quality, and relevance of learning materials in transformative ways. Large language models (LLMs) may ensure tailored and up-to-date educational materials by generating high-quality content that evolves with feedback [3]. AI further personalises learning by adapting to individual needs, such as dynamically adjusting programming exercises' complexity and providing immediate feedback [4, 5]. However, alongside this promise comes a critical challenge: ensuring the depth, reliability, and correctness of AI-generated content.

This study approaches education from the perspective of organisational needs, where agility and alignment with evolving trends are paramount. Unlike traditional academic institutions, which often retain historical materials to support foundational knowledge [6, 7], organisations must rapidly produce and update training materials to remain competitive in fast-paced industries. However, the established model for content development—where learning and development (L&D) professionals serve as intermediaries between SMEs and learners—often struggles to meet these demands. While ensuring pedagogical rigour, this process is plagued by lengthy development timelines, high costs, and outdated

AREdu 2024: 7th International Workshop on Augmented Reality in Education, May 14, 2024, Kryvyi Rih, Ukraine
olegtalaver@gmail.com (O. V. Talaver); tetianavakaliuk@gmail.com (T. A. Vakaliuk)
https://acnsci.org/vakaliuk/ (T. A. Vakaliuk)
0000-0002-6752-2175 (O. V. Talaver); 0000-0001-6825-4697 (T. A. Vakaliuk)

materials due to iterative feedback loops and disconnected workflows. Recently, many organisations have shifted to a model where SMEs create learning materials directly, bypassing the need for L&D intermediaries. While this approach allows for faster development and greater alignment with real-time organisational needs, it introduces significant challenges. SMEs, though experts in their fields, often lack instructional design expertise, which can result in unstructured or overly dense content that lacks pedagogical clarity. Additionally, SME-created materials may fail to account for diverse learner needs, leading to inconsistencies in quality and engagement. Without the support of L&D professionals, these materials may fail to provide a cohesive learning experience [8].

Recent advancements in generative AI offer a powerful alternative to L&D support. These tools can transform raw SME input into structured, learner-focused content, acting as virtual instructional designers. Despite these innovations, the risks of bias, inaccuracies, and ethical concerns remain significant. Addressing these challenges is vital to harnessing the full potential of AI in education [9, 10].

The primary objective of this research is to develop a system for processing user-provided text and flagging inaccuracies while minimising biases and ensuring the reliability of content, including AI-generated content. This objective is critical to addressing the challenges posed by inaccuracies and inconsistencies and reducing the need for extensive manual reviews.

2. Theoretical background and related works

Integrating artificial intelligence (AI) in education has reshaped traditional paradigms, introducing adaptive learning environments, enhanced accessibility, and real-time content personalisation. AI systems, such as intelligent tutoring tools and automated assessment platforms, have demonstrated the potential to foster equitable education by addressing diverse learner needs. Chiu et al. [11] emphasise the transformative impact of AI-driven adaptability in tailoring educational materials to individual learning contexts. The role of LLMs, including GPT-based systems, has been pivotal in this transformation. These models can generate dynamic course content, provide personalised instruction, and offer instant feedback, fulfilling roles such as tutors, mentors, and collaborators [12]. Despite these advancements, there remain challenges in aligning AI tools with pedagogical goals, as highlighted by Kasneci et al. [13], who warn of biases and inaccuracies in AI-generated content.

SMEs play a crucial role in enriching educational content with domain-specific knowledge, yet their contributions often lack pedagogical structure when instructional design principles are not applied [6, 8, 14]. Generative AI offers a solution by acting as a virtual instructional designer, reorganising SME inputs into structured, learner-centred materials [3]. Nazar et al. [15] illustrate how tools like CourseGPT facilitate the learning process by providing real-time, course-specific support to students, ensuring consistency and alignment with educational goals. Moreover, generative AI enhances personalisation by adapting content difficulty to learner proficiency and providing instant feedback mechanisms, as demonstrated in automated programming exercises [4]. However, the absence of oversight can lead to unstructured and ineffective content, necessitating collaborative frameworks where SMEs and AI systems work together to achieve pedagogical soundness, content richness and accuracy [3].

Adopting generative AI in education has risks, including biases in training data, content inaccuracies, and model hallucinations. These concerns are compounded by the tendency of AI systems to produce stylistically convincing but factually erroneous outputs [9, 16, 17].

Tonmoy et al. [17] have developed a comprehensive taxonomy of hallucination mitigation techniques, categorising methods into two primary domains: prompt engineering and model development. Prompt engineering encompasses methods such as retrieval-augmented generation (RAG), which integrates external knowledge sources before, during, and after generation to enhance output grounding. Specific subcategories like the Decompose-and-Query Framework and EVER Framework employ iterative validation during generation to resolve complex reasoning challenges, while post-generation tools such as RARR retrofit outputs for improved factual accuracy. Model development strategies introduce architectural innovations, such as Context-Aware Decoding and faithfulness-based loss functions, to

align AI outputs with verified truths. Additionally, supervised fine-tuning techniques incorporate counterfactual datasets and knowledge graphs to strengthen model fidelity.

Similarly, Huang et al. [9] identify core issues such as insufficient domain-specific data and over-reliance on incomplete training corpora as major contributors to AI inaccuracies. These limitations often result in hallucinations that can be categorised into factuality and faithfulness types. Factuality hallucinations occur when outputs conflict with known truths or fabricate unverifiable information, while faithfulness hallucinations involve deviations from input instructions, contextual misalignments, or logical errors. Detection approaches such as factuality, and faithfulness validation employs fact-checking, uncertainty estimation, and QA-based systems, supported by benchmarks like TruthfulQA and HaluEval, to rigorously evaluate AI outputs. Mitigation strategies span data-level interventions, such as filtering misinformation during training, to advanced inference-level techniques, including constrained sampling and logical consistency checks, thereby addressing these multifaceted challenges. Broader societal implications are explored in "Promises and challenges of generative artificial intelligence for human learning" [16], which underscores equity and ethical transparency as critical factors in the deployment of AI in education.

Considering these risks, ensuring the reliability of AI-generated materials necessitates robust, multi-layered validation frameworks. Current gaps in these processes highlight the importance of SME involvement in verifying content accuracy and contextual relevance [3]. Complementing these efforts, Shamsujjoha et al. [18] advocate for a layered runtime guardrail approach inspired by the Swiss Cheese Model, wherein distinct defensive layers tandem to intercept and correct errors at multiple stages. This model emphasises the importance of redundancy and multifaceted defences, ensuring that vulnerabilities in one layer are mitigated by the following.

Dong et al. [19] extensively reviews existing guardrail solutions, highlighting tools like Llama Guard, Nvidia NeMo, and Guardrails AI. Llama Guard demonstrates adaptability by fine-tuning specific taxonomies but depends on LLM understanding for predictive accuracy. Nvidia NeMo incorporates vector-based embeddings and KNN for content moderation and hallucination prevention. At the same time, Guardrails AI uniquely employs XML-based RAIL specifications to define output quality guarantees, including automated error correction. Despite their utility, these solutions share common weaknesses, such as lacking generalisation and holistic multi-requirement designs, which underscores the need for more robust methodologies. Dong also identifies technical challenges, particularly unintended responses, which are detected through adversarial prompt testing and mitigated via adversarial training, safety reinforcement, and input/output rephrasing. Fairness remains a critical issue, with biases categorised into gender, cultural, and dataset origins. Protective measures, including fine-tuning and fairness-driven prompt engineering, are essential but require continuous learning and diversification of datasets. Privacy concerns, such as PII leakage, are addressed through differential privacy methods and personalised watermarking to ensure ownership and confidentiality. Hallucinations pose significant risks; detection strategies like self-consistency checks and external validations are complemented by protective measures such as retrieval-augmented generation and Verify-and-Edit frameworks [20]. These approaches balance robustness and adaptability, integrating rigorous software development life cycles and Pareto optimisation [19] for practical guardrail implementations.

3. Methods

The system was developed to process user-provided text, perform factual verification, and generate revisions where inconsistencies were identified. Building on established methodologies [15, 17, 20], the workflow incorporated multi-stage processing, structured claim extraction, evidence classification, and revision strategies. Prompt engineering techniques were employed, including prompt tuning with prompt chaining and multiple stages of results refinement. Most importantly, verification occurs after generation, meaning that the system receives and analyses any text. Decisions are made based on the retrieved information from various sources, including vectorised documents. The design emphasised adherence to factual accuracy while minimising stylistic alterations.

Testing was conducted using articles that sometimes included altered factual information to evaluate the system's ability to detect inaccuracies. Purely opinion-driven texts were excluded to ensure the dataset's alignment with the system's verification scope.

The input text is segmented into paragraphs and sentences. Next, claims are extracted from each sentence and then refined on the whole paragraph level to remove duplicates and improve the context for already-created claims.

Claims extraction prompt:

```
Extract all factual claims from the given text. A factual claim is any
statement that asserts something as a fact and can be verified.
Guidelines:
1. The 'claim_text' should distill the factual assertion while preserving
its meaning. Remove unrelated phrases or context.
2. The 'original_text' must exactly match the corresponding part of the
input text from which the claim was derived.
3. Each claim must contain only 1 factual information and enough context to
be unambiguous and easy to search, with context information, specific dates,
names, and people etc.
4. The claim must question single fact, not multiple facts, so that it's
easy to search but have enough context to be unambiguous.
6. Select a particular part of the original text when referencing it, not
the entire sentence
7. Do not include opinions, speculative statements, or rhetorical questions
as claims.
8. If there is ambiguity in the claim, provide the most concise and accurate
interpretation.
9. Ensure claims are extracted accurately, even if they are embedded in
longer sentences.
```

Then, the system retrieves evidence from external Google using Google Custom Search, Wikipedia, and a vector store for each claim. Evidence snippets are individually classified as Supporting, Contradicting, or Neutral based on relevance, which the model itself determines. Also, the model is instructed to output a credibility score based on the website on which the material was found. Irrelevant or low credibility are filtered out. Classification utilises chain-of-thought (CoT) reasoning, wherein the LLM processes claims and evidence incrementally to facilitate nuanced judgments.

Evidence analysis prompt:

```
Evaluate the relationship between the given claim and the provided evidence.
For each claim, determine whether the evidence supports, contradicts, or is
neutral to the claim.
Criteria for Decision:
- {DecisionClaimEnum.supporting}: Evidence explicitly affirms the claim
with no ambiguity.
- {DecisionClaimEnum.contradicting}: Evidence explicitly denies or disputes
the claim with no ambiguity.
- {DecisionClaimEnum.neutral}: Evidence is unrelated, ambiguous, or
insufficient to affirm or deny the claim.
Additionally, assess the credibility of the source using the following
scale:
- 1.0: Peer-reviewed resources (e.g., white papers, academic journals)
- 0.8: Crowd-sourced resources with robust moderation (e.g., Wikipedia)
- 0.6: Reputable news outlets and well-known publications
```

```
- Below 0.6: Sources with questionable credibility or lack of verification
The result should include:
1. Decision: '{DecisionClaimEnum.supporting}', '{DecisionClaimEnum.
contradicting}', or '{DecisionClaimEnum.neutral}'
2. Source Credibility: A float value between 0 and 1, as defined above
3. Extract key phrases from the evidence that support the decision.
4. Provide a brief justification for the credibility score.
Additional Step:
- Relevance: Assess whether the evidence is contextually relevant to the
claim before determining the relationship.
- If evidence is irrelevant, set 'relevant' field to false and the decision
should default to '{DecisionClaimEnum.neutral}'
```

Classifications for each claim were aggregated by counting instances of supporting, contradicting, and neutral evidence. A weighted ratio method was employed to assess the certainty of each claim's correctness or incorrectness based on the credibility of evidence classified as supporting or contradicting. Each piece of evidence is assigned a credibility score (w_i) within a range of 0 to 1, reflecting the source's reliability (see evidence analysis prompt above). Certainty calculations consider the aggregated credibility and the total number of supporting (n_s) and contradicting (n_c) evidence items. The certainty that a claim is incorrect is calculated using the formula:

$$CertaintyIncorrect = \frac{ConTotal}{SupTotal + ConTotal} \tag{1}$$

where $SupTotal = \sum_{i=1}^{n_s} w_i$, the aggregated credibility of all supporting evidence; $ConTotal = \sum_{j=1}^{n_c} w_j$, the aggregated credibility of all contradicting evidence.

This ratio reflects the proportion of contradicting evidence relative to the total weighted evidence (supporting and contradicting). This score is compared against a predefined threshold (0.6 in this case) to determine if the claim should be flagged for revision if we are certain enough that it is incorrect.

When no supporting or contradicting evidence is provided ($n_s = 0$, $n_c = 0$), CertaintyIncorrect is set to 0, as the absence of evidence does not permit any judgment on correctness. Claims are not flagged for revision in this case.

When only supporting evidence is provided ($n_s > 0$, $n_c = 0$), certainty is determined by the average credibility of the supporting evidence:

$$CertaintyCorrect = \frac{SupTotal}{n_s} \tag{2}$$

If the average supporting credibility exceeds a threshold, the claim is considered correct and left unchanged; otherwise, it's considered uncertain. This does not trigger revisions due to the lack of contradicting evidence, which is the source of the revised version of the paragraph.

When only contradicting evidence is provided ($n_s = 0$, $n_c > 0$), the certainty that the claim is incorrect is determined by the average credibility of the contradicting evidence, similar to the previous case:

$$CertaintyIncorrect = \frac{ConTotal}{n_c} \tag{3}$$

However, this time, we also check if the threshold has not exceeded (0.6); if it does, we are confident enough to revise the paragraph text based on the contradicting evidence (use those as a source of truth).

It must be noted that this methodology filters out neutral evidence, which may lead to the loss of nuanced information. Incorporating a weighting factor for neutral evidence could improve robustness. Also, evidence with extremely high or low credibility scores (e.g., 1.0 or near 0) can disproportionately skew the results. Implementing a credibility cap or floor may mitigate this effect. Finally, the method switches from ratio-based certainty calculations (for mixed evidence) to average-based certainty in

edge cases (only supporting or contradicting), which can lead to perceived inconsistency in certainty interpretation.

4. Results

A simple web UI was developed to support better visualisation of results (the full listing is available here: https://github.com/Vidzhel/ai-fact-checker). The page has a field where the provided text is placed. Upon clicking "Verify", the whole text is sent for verification, which takes approximately 1 minute for up to 5 paragraphs of text (figure 1).

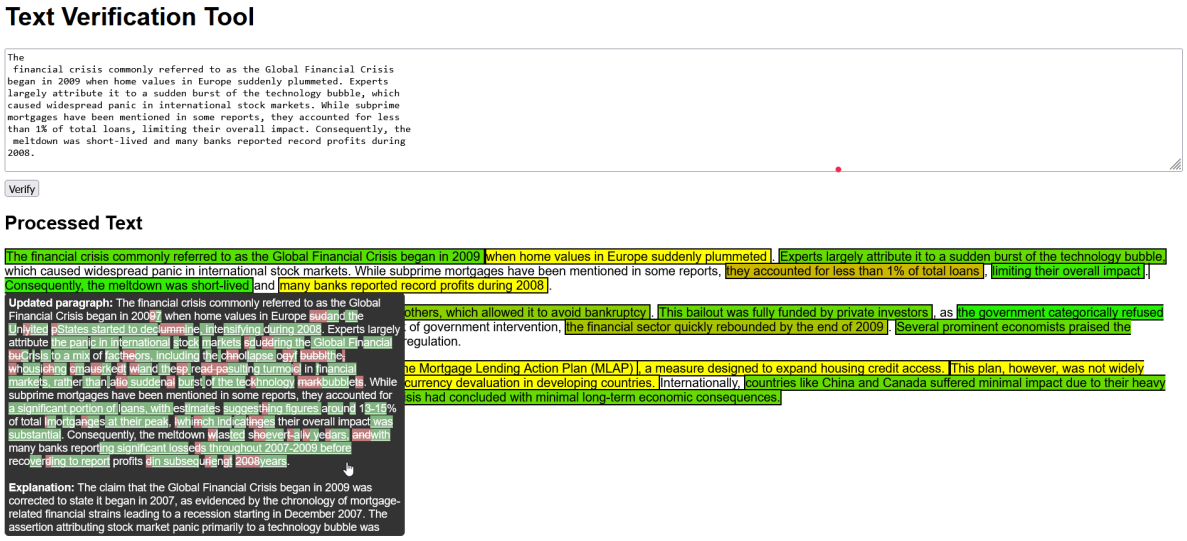

Figure 1: An example of the verification tool UI includes a field with inserted text and processed output that provides suggestions for different parts and a popup with details that are opened upon hovering highlighted parts of the text.

The processed text contains the same paragraphs with highlighted parts representing analysed parts of the text – claims. The colour of highlighting varies depending on the certainty of the decision (despite the paragraph being changed or not), where yellow indicates that the provided evidence seems to support the claim, and green indicates that the evidence is very likely to support the claim (basically gradient based certainty level ranging from 0.5 to 1). If the certainty level is below 0.5 – the highlighting will be red, indicating uncertainty. The highlighted parts can have either a black or blue border colour, indicating that the piece was or wasn't edited respectively.

Upon hovering over a highlighted part, a popup shows all the details related to the claim. Revised paragraphs start with the difference between original and revised paragraphs, and an explanation of the changes is summarised below. Then, there are details related to the certainty of the decision – the score and aggregated evidence counts. After that, a list of evidence is specified, each having the link, credibility level (with an explanation of the decision), a snippet of text used, a summary of findings related to the claim and classification (supporting, neutral, contradicting). This helps the reviewer make an informed decision [21].

Shortened example of details shown upon hovering over the highlighted part that was rewritten:

```
Updated paragraph: The financial crisis commonly referred to as the Global
Financial Crisis began in 2007 when home values in Europe  and the United
States started to decline, intensifying during 2008. Experts largely attribute
the panic in international stock markets during the Global Financial Crisis
to a mix of factors, including the collapse of the housing market and the
```

```
resulting turmoil in  financial markets, rather than a sudden burst of the
technology bubble. While subprime mortgages have been mentioned in some
reports, they accounted for a significant portion of loans, with estimates
suggesting figures around 13-15% of total mortgages at their peak, which
indicates their overall impact was substantial. Consequently,  the meltdown
lasted several years, with many banks reporting significant losses throughout
2007-2009 before recovering to report profits in subsequent years.

Explanation: The claim that the Global Financial Crisis began in 2009 was
corrected to state it began in 2007, as evidenced by the chronology of mortgage-
related financial strains leading to a recession starting in December 2007. The
assertion attributing stock market panic primarily to a technology bubble was
revised to reflect the documented role of the housing market collapse and broader
financial crisis. The claim about subprime mortgages being less than 1% was
adjusted to acknowledge their substantial impact, aligning with evidence that
they represented a significant portion of lending. Lastly, the characterization
of the crisis as short-lived was corrected to acknowledge its prolonged effects,
lasting from 2007 into 2009.

Certainty: 0.87 - Based on gathered data - decision seems to be very likely
correct (whether claim was altered or not);
Supporting: 6 - Contradicting: 0 - Neutral: 1
Source: https://www.federalreservehistory.org/essays/subprime-mortgage-crisis
Cridability: 1 - The source is a reputable, peer-reviewed resource from the
Federal Reserve, which provides credible historical context and analysis.
Snippet: The expansion of mortgages to high-risk borrowers, coupled with
rising house prices, contributed to a period of turmoil in financial
markets that lasted from ...
Evidence: expansion of mortgages to high-risk borrowers...contributed to
a period of turmoil in financial markets
Classification: Supporting

Source: https://en.wikipedia.org/wiki/2007%E2%80%932008_financial_crisis
Cridability: 0.8 - The source is Wikipedia, a crowd-sourced resource with
robust moderation, making it a credible source for general information.
Snippet: The 2007–2008 financial crisis, or the global financial crisis
(GFC), was the most severe worldwide economic crisis since the 1929 Wall
Street crash that began the Great Depression. Causes of the crisis
included predatory lending in the form of subprime mortgages to low-income
homebuyers and a resulting ... in the 2010s European debt crisis.
Evidence: Causes of the crisis included predatory lending in the form of
subprime mortgages to low-income homebuyers
Classification: Supporting
```

5. Discussion

The outlined method offers a clear and practical framework for extracting, verifying, and refining claims, prioritising simplicity and effectiveness. Combining prompt engineering with a multi-stage evidence analysis process, the system consistently identifies claims within the text, retrieves relevant evidence, and refines results, making the review process more straightforward. Using prompt-chaining techniques, the system divides verification into smaller, manageable steps, such as clarifying ambiguous references

to entities, dates, or locations before retrieving evidence. This iterative approach incrementally improves accuracy and minimises errors caused by incomplete context.

While the system performs well in detecting factual inaccuracies, it faces some notable challenges. These include occasionally missing specific claims within a passage and failing to provide adequate context for specific claims. For instance, a statement like "the crisis significantly impacted the current economic situation" may lack explicit reference to timeframes or events, preventing effective retrieval of evidence. Resolving vagueness or insufficient detail issues is crucial for improving the system's reliability.

Addressing these shortcomings could involve fine-tuning the model with more varied and representative training examples. This would enhance its ability to recognise complex claims, define contextual boundaries, and retrieve evidence more precisely. Additionally, using vector embeddings, such as those from the text-embedding-3-small model, significantly boosts the system's capability for accurate evidence retrieval, particularly in specialised fields requiring alignment with specific terminology and context.

Another strength of the approach is its cost efficiency. By using a combination of GPT-4o-mini for generating completions and text-embedding-3-small for retrieval, the system processes approximately 3,000 pages (two million tokens) for just $0.50. Including a structured API further ensures robust performance by delivering machine-readable outputs validated against schemas. This consistency simplifies subsequent verification and refinement while reducing the risk of errors in output formatting.

6. Conclusions

This study presents a efficient approach for processing and verifying user-provided text, focusing on fact-checking provided text with evidence extracted from various sources. The system balances simplicity and performance by integrating embeddings for domain-specific retrieval and prompt-engineered reasoning techniques, making it a promising tool for real-world applications. The proposed system effectively extracts claims, retrieves relevant evidence, and classifies evidence to refine text based on credibility-weighted thresholds, allowing it to either prove fact correctness or mark and suggest an update for inaccurate text parts.

Despite these strengths, limitations remain. The system occasionally struggles to handle ambiguous claims or claims lacking sufficient context, complicating evidence retrieval and verification. Addressing these issues through fine-tuning can significantly enhance the system's contextual understanding and precision. Also, including more test data would be beneficial.

Compared to similar tools, this system stands out for its emphasis on combining affordability with a structured approach to ensure consistent results. Unlike generic models prioritising breadth over domain specificity, the embedding-based retrieval method employed here enhances precision, particularly in specialised fields.

Overall, this study highlights the potential of combining simple yet powerful techniques to address text verification and content refinement challenges. The findings demonstrate that a structured, scalable approach can provide reliable results while opening avenues for further refinement and adoption in diverse applications. Researchers and developers are encouraged to explore and refine this system in real-world settings, ensuring its full potential is realised across domains.

Declaration on Generative AI: During the preparation of this work, the authors used GPT-4o in order to: Generate literature review, Abstract drafting, Content enhancement. After using this service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. V. Marienko, S. O. Semerikov, O. M. Markova, Artificial intelligence literacy in secondary education: methodological approaches and challenges, CEUR Workshop Proceedings 3679 (2024) 87–97.

[2] I. Mintii, S. Semerikov, Optimizing Teacher Training and Retraining for the Age of AI-Powered Personalized Learning: A Bibliometric Analysis, in: E. Faure, Y. Tryus, T. Vartiainen, O. Danchenko, M. Bondarenko, C. Bazilo, G. Zaspa (Eds.), Information Technology for Education, Science, and Technics, volume 222 of *Lecture Notes on Data Engineering and Communications Technologies*, Springer Nature Switzerland, Cham, 2024, pp. 339–357. doi:10.1007/978-3-031-71804-5_23.

[3] D. Leiker, S. Finnigan, A. R. Gyllen, M. Cukurova, Prototyping the use of Large Language Models (LLMs) for Adult Learning Content Creation at Scale, in: S. Moore, J. C. Stamper, R. J. Tong, C. Cao, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh, C. Brooks (Eds.), Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023), Tokyo, Japan, July 7, 2023, volume 3487 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 3–7. URL: https://ceur-ws.org/Vol-3487/short1.pdf.

[4] Y. Bauer, J. P. Leal, R. Queirós, Authoring Programming Exercises for Automated Assessment Assisted by Generative AI, in: A. L. Santos, M. Pinto-Albuquerque (Eds.), 5th International Computer Programming Education Conference (ICPEC 2024), volume 122 of *Open Access Series in Informatics (OASIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2024, pp. 21:1–21:8. doi:10.4230/OASIcs.ICPEC.2024.21.

[5] E. Dickey, A. Bejarano, C. Garg, AI-Lab: A Framework for Introducing Generative Artificial Intelligence Tools in Computer Programming Courses, SN Computer Science 5 (2024) 720. doi:10.1007/s42979-024-03074-y.

[6] C. Zuo, Research on the Training of College Teachers' Practical Teaching Ability under the Background of the Application-Oriented Transformation, in: Proceedings of the 2017 7th International Conference on Education and Management (ICEM 2017), Atlantis Press, 2018, pp. 275–278. doi:10.2991/icem-17.2018.58.

[7] H. Wang, Research on the Development of College Teachers' Practical Ability from the Perspective of Teachers' Professional Development, in: Proceedings of the 2016 International Conference on Management Science and Innovative Education, Atlantis Press, 2016, pp. 544–549. doi:10.2991/msie-16.2016.120.

[8] K. Spiro, V. Bhamidi, Employee-Generated Learning: How to Develop Training That Drives Performance, Kogan Page, 2024. URL: https://books.google.com.ua/books?id=dwChzwEACAAJ.

[9] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43 (2025) 42. doi:10.1145/3703155.

[10] S. Ayyamperumal, L. Ge, Current state of LLM Risks and AI Guardrails, 2024. doi:10.48550/arXiv.2406.12934.

[11] T. K. F. Chiu, Q. Xia, X. Zhou, C. S. Chai, M. Cheng, Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education, Computers and Education: Artificial Intelligence 4 (2023) 100118. doi:10.1016/j.caeai.2022.100118.

[12] E. Mollick, L. Mollick, Assigning AI: Seven Approaches for Students, with Prompts, 2023. doi:10.2139/ssrn.4475995.

[13] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274. doi:10.1016/j.lindif.2023.102274.

[14] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, Z. Du, Artificial intelligence in education: A systematic literature review, Expert Systems with Applications 252 (2024) 124167. doi:10.1016/j.eswa.2024.124167.

[15] A. M. Nazar, M. Y. Selim, A. Gaffar, S. Ahmed, Revolutionizing Undergraduate Learning: CourseGPT and Its Generative AI Advancements, 2024. URL: https://arxiv.org/abs/2407.18310. arXiv:2407.18310.

[16] L. Yan, S. Greiff, Z. Teuber, D. Gasevic, Promises and challenges of generative artificial in-

telligence for human learning, Nature Human Behaviour 8 (2024) 1839–1850. doi:10.1038/s41562-024-02004-5.

[17] S. M. T. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, A. Das, A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, 2024. URL: https://arxiv.org/abs/2401.01313. arXiv:2401.01313.

[18] M. Shamsujjoha, Q. Lu, D. Zhao, L. Zhu, Designing Multi-layered Runtime Guardrails for Foundation Model Based Agents: Swiss Cheese Model for AI Safety by Design, 2024. URL: https://arxiv.org/abs/2408.02205. arXiv:2408.02205.

[19] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, X. Huang, Position: building guardrails for large language models requires systematic design, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024, p. 451. URL: https://dl.acm.org/doi/10.5555/3692070.3692521.

[20] R. Zhao, X. Li, S. Joty, C. Qin, L. Bing, Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5823–5840. doi:10.18653/v1/2023.acl-long.320.

[21] N. Ishizu, W. L. Yeoh, H. Okumura, O. Fukuda, The Effect of Communicating AI Confidence on Human Decision Making When Performing a Binary Decision Task, Applied Sciences 14 (2024) 7192. doi:10.3390/app14167192.