EduCheck: A Benchmark and Supervised Fine-Tuning Approach for Detecting Pedagogical and Conceptual Hallucinations in AI-Generated Educational Content

1st Jun-Wei Zeng*
Vanke Meisha Academy
Shenzhen, China
zengjunwei@stu.vma.edu.cn
0009-0006-1920-7118

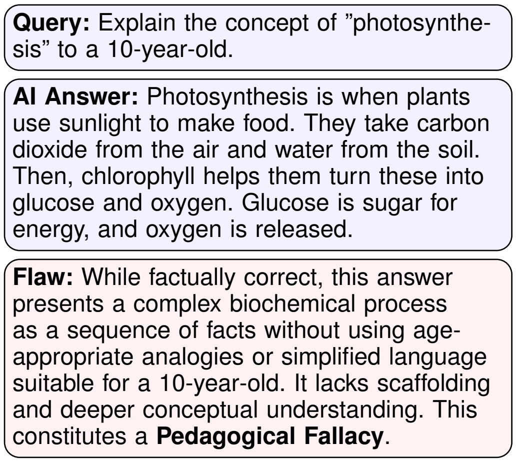
***Abstract*—With the rapid development of Large language models (LLMs) in educational fields, assurance of accurate and pedagogical content is of essential importance. Most existing papers on hallucination detection primarily focus on factual errors or general facts, overlooking deeper flaws in educational content: Conceptual misinterpretations and pedagogical Fallacies. To address this research gap, our work proposes a novel taxonomy specifically for hallucinations in educational content. Based on this taxonomy, we constructed EduHallucinate, a dataset focused on these deep hallucinations, comprising more than 800 tagged samples. Finally, utilizing supervised fine-tuning(SFT) combined with a weighted loss function, we trained our detection system, EduCheck-SFT, on a lightweight pre-trained model, distilbert-base-uncased. Experients results show that our model's robustness, achieving a F1-score of 75.4% and a Recall of 100% on a test set of 140 samples. This shows the effectiveness of our approach in identifying pedagogical flaws that were not addressed by former research.**

***Index Terms*—Hallucination Detection, AI4Education, LLMs, SFT, NLP**

I. Introduction

Artificial intelligence, particularly Large Language Models(LLMs), is rapidly changing the field of education. They provide support in certain areas: personalized tutoring, automated content generation and instant question answering [1]. However, hallucination occurs, generating plausible but factually incorrect or unreliable content. In the high-stakes domain of education, content accuracy is non-negotiable [2]. A flawed concept may lead to long-lasting, hard-to-correct damage to students' knowledge base.

While existing research on hallucination detection has made notable progress, its primary focus has been on internal reasons of hallucination occurrence and general factual error detection [3], [4]. Based on former research, specific area studies are suggested, and general research is insufficient for educational scenarios [3]. An AI response that is factually correct but pedagogically flawed can be more detrimental than a simple factual inaccuracy. For instance, a problem-solving process that skips crucial steps and encourages simple memorization constitutes a harmful Pedagogical Fallacy. Similarly, a Conceptual Misinterpretation using an inappropriate analogy can embed an incorrect cognitive model in a student's mind. These ”deep hallucinations” are often invisible to current general-purpose detection systems.



Query: Explain the concept of ”photosynthesis” to a 10-year-old.

AI Answer: Photosynthesis is when plants use sunlight to make food. They take carbon dioxide from the air and water from the soil. Then, chlorophyll helps them turn these into glucose and oxygen. Glucose is sugar for energy, and oxygen is released.

Flaw: While factually correct, this answer presents a complex biochemical process as a sequence of facts without using age-appropriate analogies or simplified language suitable for a 10-year-old. It lacks scaffolding and deeper conceptual understanding. This constitutes a **Pedagogical Fallacy**.

Figure 1: An example of a pedagogically flawed AI-generated educational content. The AI's answer, while factually accurate, fails to apply appropriate pedagogical strategies for the target audience.

To fill the critical gap, this paper makes the following three contributions:

1) We systematically define and propose a novel taxonomy for hallucinations specifically tailored to AI-generated educational content, extending beyond factual correctness to include pedagogical effectiveness.Including **Factual Error**, **Conceptual Misinterpretation** and **Pedagogical Fallacy**.
2) We constructed **EduHallucinate**, a high-quality, annotated benchmark dataset of over 800 samples, focusing on identifying factual errors, conceptual misinterpreta-

Identify applicable funding agency here. If none, delete this.

tions, and pedagogical fallacies. To our knowledge, this is the first dataset dedicated to these types of deep hallucinations in education.

3) We successfully trained a detection model, **EduCheck-SFT**, demonstrating that even a lightweight model, when fine-tuned on our specialized dataset, can effectively identify these complex, previously overlooked hallucination types, achieving a safety-prioritized high-recall performance.

Our detection system, **EduCheck-SFT**, utilizes supervised fine-tuning on the lightweight **distilbert-base-uncased** Transformer architecture, specifically trained on our **EduHallucinate dataset** to recognize the unique patterns associated with these deep educational hallucinations.

II. Related Works

To position our work, we review literature across three key areas. First, we examine existing taxonomies and detection methods for hallucinations in Large Language Models (LLMs), highlighting their typical focus. Second, we survey benchmark datasets currently used for evaluating LLM hallucinations, noting their scope and limitations. Finally, we discuss the context of AI evaluation within the educational domain (AIEd), contrasting existing approaches with our focus on intrinsic content quality. This review will demonstrate the specific gap addressed by our research: the lack of methods and resources dedicated to detecting conceptual and pedagogical flaws in AI-generated educational content.

A. Hallucination Taxonomies and Detection Methods

The phenomenon of ”hallucination” in LLMs is defined as the generation of content that is plausible yet factually incorrect, nonsensical, or unsubstantiated by provided source material. As a result, it is widely seen as a critical barrier that hinders the reliable application of LLMs across various fields [9].

Research on LLM hallucinations can be broadly categorized into two streams [9]. The first investigates the internal mechanism and causation of hallucinations, such as knowledge representation or training data artifacts, aiming to mitigate them at the source. The second stream, which our work contributes to, focuses on the external, post-hoc detection of hallucinations. This research aims to build practical systems that can validate generated content after it has been generated.

Within the second stream of external detection, a review of the literature reveals that existing research predominately classifies hallucinations along two primary axes: factuality and faithfulness [10]. Factual hallucinations refer to output that contradict verifiable real-world knowledge. Faithfulness hallucinations, are more common in summarization or Retrieval-Augmented Generation(RAG), occur when the output differ from a source context [9].

Consequently, current detection methods are primarily designed to identify these specific error types. Methodologies range from binary classifiers and uncertainty-based detection to more advanced approaches. For instance, Su et al.(2025) [3] employed reinforcement learning (RL) to achieve fine-grained span detection of hallucinated content.

However, these methods, primarily designed for factuality or faithfulness verification against a source, are ill-equipped to evaluate the intrinsic instructional quality demanded in education. They do not address the novel categories of errors crucial for learning outcomes—Conceptual Misinterpretations and Pedagogical Fallacies—which forms the central motivation for our work.

Our work complements this body of research by focusing on a more specialized, high-stakes domain: education. We address a novel category of errors crucial for learning outcomes: Conceptual Misinterpretation and Pedagogical Fallacy.

B. Benchmark Datasets for Hallucination Evaluation

The field’s focus on factuality and faithfulness is directly reflected in current benchmark datasets. For instance, TruthfulQA [11] is designed to measure a LLM’s propensity to mimic human falsehoods, with annotations focused entirely on factual correctness. Other datasets, such as HaluEval [12], are constructed to evaluate faithfulness by checking for fabrication that is not presented in the original source.

While these datasets are essential for general-purpose models, they are insufficient for the educational domain. These datasets do not allow for the detection of the ”deep hallucination” we propose. For example, no existing benchmark is annotated to identify when an AI uses a flawed analogy (a Conceptual Misinterpretation) or provides an educational output(a Pedagogical Fallacy). This highlights a critical gap: the lack of a specialized, expert-annotated benchmark dataset focused specifically on the conceptual integrity and pedagogical soundness of AI-generated educational content, a gap our EduHallucinate dataset aims to fill.

C. Evaluating AI in Education (AIEd)

The integration of AI, and LLMs in particular into education settings is a rapidly expanding field of research [1], [2]. Studies actively explore the use of Generative AI in specialized domains like coding education [4], its role in student problem solving in subjects like chemistry [7], and its general performance as a science educational tool [8]. Kasneci et al.(2023) [2] provide a comprehensive overview of both the advancements and significant challenges that AI brought in education.

However, existing research on the evaluation of these AIEd tools often centers on their downstream impact. For instance, Almasri (2024) [1] provides a systematic review that illustrate AI’s impact on student engagement and learning outcomes. This focus on downstream is crucial, but it overlooks the intrinsic quality and pedagogical safety of AI-generated content itself.

Effective teaching or pedagogy is a complex discipline that required nuanced strategies, such as adapting content to student learning styles [6] or utilizing constructivist learning approaches [5]. Pedagogical Fallacy may occur when an AI tool generates factually correct but ignores learning styles [6] or oversimplifying answers instead of scaffolding problem-solving [7]. While studies have begun to explore AI's effect on conceptual knowledge [7], [8], a systematic method for detection of pedagogical and conceptual flaws are crucial and demanded to address the research gap. See Table 1.

Application Domain	Factual / Faithfulness Errors	Conceptual / Pedagogical Errors
Educational Domain	*(Relatively Unexplored regarding deep pedagogical issues)*	**EduCheck (This Work)** Focuses on Conceptual Misinterpretations and Pedagogical Fallacies specific to educational content.
General Domain	TruthfulQA, HaluEval, Su et al. (RL4HS) Focus on factual correctness and faithfulness to source in general texts.	*(Typically not the focus in general domain detection.)*

TABLE 1: Positioning EduCheck within the landscape of hallucination detection research based on application domain and error type focus. Existing work primarily targets factual/faithfulness errors in general domains, while EduCheck addresses the novel area of conceptual/pedagogical errors within education.

While studies have begun to explore AI's effect on conceptual knowledge [7], [8], a systematic methodology and dedicated benchmark for detecting intrinsic pedagogical and conceptual flaws within AI-generated content itself remain absent. Addressing this specific research gap is the primary objective of the EduCheck project.

III. Methodology

A. A Taxonomy of Educational Hallucinations

To systematically study this problem, we propose the following three categories of hallucinations relevant to educational content, see Figure 2:

1) **Factual Error**: The content contains explicit information contradicting established scientific facts, formulas, or historical events.

   Example (Chemistry): "The Bohr Model is the current, accepted model of the atom." *(Incorrect: The Bohr model is an outdated historical model).*

2) **Conceptual Misinterpretation**: The content, while potentially factually accurate at a surface level, uses inappropriate analogies, flawed causal reasoning, or ambiguous descriptions that lead the student to form an incorrect cognitive model of an abstract concept.

   Example (Computer Science): "Encapsulation...is primarily done to allow the user to modify an object's internal attributes directly." *(Incorrect: Encapsulation aims to prevent direct access).*

3) **Pedagogical Fallacy**: The content itself might be factually correct, but its presentation, explanation, or the guided process violates fundamental pedagogical principles, potentially encouraging rote memorization, hindering critical thinking, or promoting bad practices.

   Example (Biology): "The mitochondria is the powerhouse of the cell." *(Pedagogical Fallacy: This is factually correct but is a classic example of encouraging rote memorization. It provides no conceptual understanding of how mitochondria function, e.g., through cellular respiration and ATP synthesis.)*

This taxonomy provides the analytical framework for the construction of our benchmark dataset, EduHallucinate, which we detail in the following section.

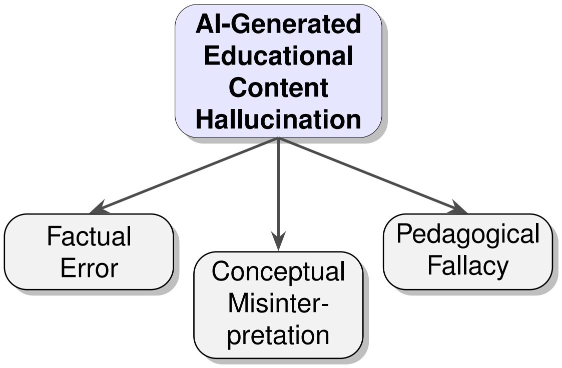

AI-Generated Educational Content Hallucination
Factual Error
Conceptual Misinterpretation
Pedagogical Fallacy

Figure 2: The proposed taxonomy of hallucinations in AI-generated educational content, categorizing flaws into Factual Errors, Conceptual Misinterpretations, and Pedagogical Fallacies.

B. The EduHallucinate Benchmark Dataset

We constructed the EduHallucinate dataset through the following process:

1) **Domain Selection**: We chose high-school level Chemistry (a fact-based hard science) and high-school level Computer Science (a logic-based engineering science) to ensure diversity and applicability across different educational domains.
2) **Content Generation**: We designed a structured prompt template instructing an LLM (Gemini 2.5 flash, GPT-5) to act as a high school teacher and generate explanatory texts for a curated list of "hard-core" concepts known to be prone to errors or misconceptions.

   An example of the prompt template used is as follows:

   Prompt: You are an expert high school [Subject] teacher (e.g., Chemistry, Computer science) with 10 years of experience. You are known for making complex topics easy to understand.

   Task: Generate a clear, concise, and accurate explanation for the following high school-level concept:

Concept: {concept_name}

Constraints:

a) Your explanation must be factually correct.
b) **CRITICAL:** Avoid common misconceptions. For {concept_name}, a common misconception is {known_misconception}. You must *not* reinforce this.
c) Use clear analogies, but ensure they are conceptually sound.
d) Do not just state a rule; briefly explain *why* it is true.

This template was then programmatically filled with concepts such as ”Thermodynamics” or ”loop” and their associated common misconceptions.

3) **Expert Annotation**: All generated samples were meticulously reviewed and annotated by the author and experts. Each sample was labeled as containing a hallucination ('Y') or not ('N'). For hallucinated samples, the specific type (Factual, Conceptual, Pedagogical) was identified, and detailed notes explaining the error and suggesting corrections were provided. Each generated text sample x was manually reviewed and assigned a binary label $y \in \{0,1\}$, where $y = 1$ indicates the presence of any hallucination (Factual, Conceptual, or Pedagogical) and $y = 0$ indicates clear content. Detailed notes justifying the label and identifying the specific problematic span were recorded for samples where $y = 1$.
4) **Dataset Statistics**: The final version of EduHallucinate contains 839 samples, divided into a training set of 697 samples and a test set of 142 samples. Approximately 58% of the samples in the dataset are labeled as containing a hallucination.

TABLE 2: Distribution of the EduHallucinate Test Set (N=142)

Category	**Count**	**Percentage**
Total Hallucination ($y = 1$)	86	60.6%
- Factual Error (FE)	35	24.6%
- Conceptual Misinterpretation (CM)	29	20.4%
- Pedagogical Fallacy (PF)	22	15.5%
No Hallucination ($y = 0$)	56	39.4%

Note: The proportions (60% vs 40%) demonstrate the emphasis on errors within the dataset.

C. The EduCheck-SFT Detection Model

To validate the effectiveness of our dataset, we trained a binary classification model using Supervised Fine-Tuning (SFT).

1) **Base Model**: We selected distilbert-base-uncased, a lightweight yet efficient Transformer model.
2) **Training Strategy**: Prioritizing safety in the educational context (”better safe than sorry”), we aimed to maximize the model's recall. To achieve this, we employed a Weighted Loss Function during training, calculating class weights ([1.1976, 0.8584]) to impose a heavier penalty for misclassifying hallucinated samples (false negatives). To explicitly prioritize the detection of hallucinated content (the minority class, $y = 1$), we employed a weighted cross-entropy loss function during fine-tuning. The loss L is computed as:

$$L = -\sum_i (w_1 y_i \log(\hat{y}_i) + w_0(1 - y_i)\log(1 - \hat{y}_i)) \quad (1)$$

where y_i is the true label for sample i, $\hat{y}_i$ is the model's predicted probability for the positive class, and w_0, w_1 are the class weights computed as inversely proportional to the class frequencies in the training set ($w_0 \approx 0.86$, $w_1 \approx 1.20$). This imposes a higher penalty for misclassifying the critical hallucination samples.
3) **Early Stopping**: To mitigate overfitting on our dataset, we implemented an early stopping strategy. Training was automatically halted when the F1-score on the evaluation set (test set) did not improve for 2 consecutive epochs, and the best-performing model checkpoint was saved (achieved at epoch 3).

While supervised fine-tuning is a standard technique, the primary originality of our methodology lies in its application to the novel task defined by our taxonomy and enabled by our unique EduHallucinate dataset. We are, to our knowledge, the first to specifically train a model to detect conceptual misinterpretations and pedagogical fallacies in AI-generated educational text, demonstrating the feasibility of identifying these subtle but critical errors beyond simple fact-checking.

IV. Experiments and Results

A. Experimental Setup

All experiments were conducted locally on a MacBook Pro equipped with an Apple M4 Pro chip, utilizing the PyTorch framework (version 2.2.2) with its Metal Performance Shaders (MPS) backend for GPU acceleration. We used the Hugging Face Transformers library (version 4.57.1)

Dataset: We used the EduHallucinate dataset detailed in Section 3.2, comprising 839 samples. The dataset was split into a training set of 697 samples and a test set of 142 samples. Within the test set, 86 samples were labeled as containing hallucinations ($y = 1$) and 56 samples were labeled as clear ($y = 0$).

Base Model: We employed distilbert-base-uncased [13] as our base pre-trained language model, chosen for its balance between performance and computational efficiency.

Training Details: The model was fine-tuned for a maximum of 15 epochs using the AdamW optimizer [14] with a learning rate of 1×10^{-5} and a weight decay of 0.01. We used a per-device batch size of 1 and gradient accumulation steps of 16 (effective batch size of 16). The weighted cross-entropy loss function described in Section 3.3 was used. An early stopping strategy was implemented with a patience of 2 epochs, monitoring the F1-score on the test set, which resulted

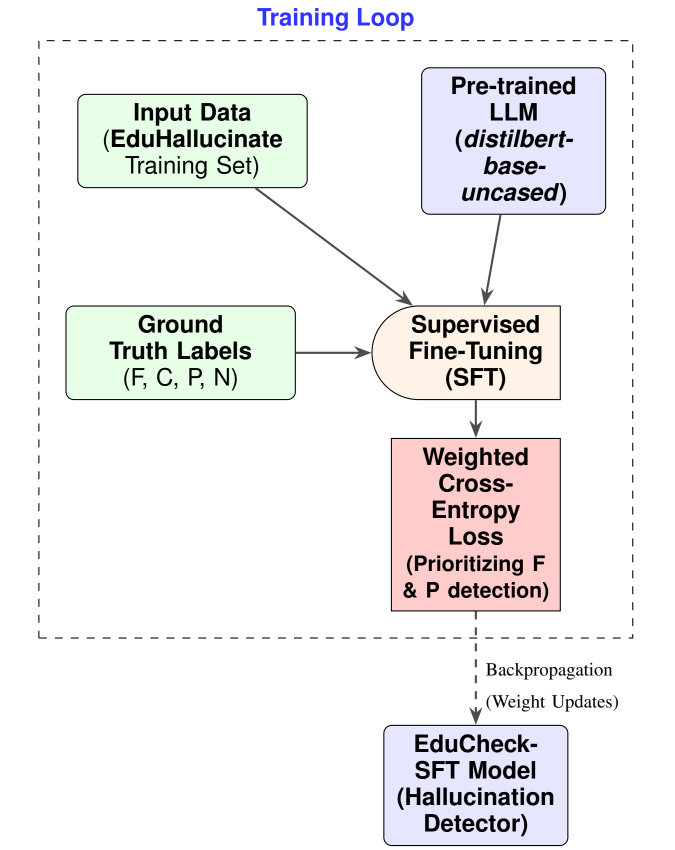

Training Loop
Input Data (EduHallucinate Training Set)
Pre-trained LLM (*distilbert-base-uncased*)
Ground Truth Labels (F, C, P, N)
Supervised Fine-Tuning (SFT)
Weighted Cross-Entropy Loss (Prioritizing F & P detection)
Backpropagation (Weight Updates)
EduCheck-SFT Model (Hallucination Detector)

Figure 3: Architecture of the Supervised Fine-Tuning (SFT) process for the EduCheck-SFT model. The process utilizes the EduHallucinate dataset and incorporates a specialized weighted loss function to prioritize the detection of critical Factual and Pedagogical errors.

in the training stopping after epoch 3 and loading the best performing model from epoch 1.

Evaluation Metrics: Model performance was evaluated using standard classification metrics: Accuracy, Precision, Recall, and F1-Score, calculated for the positive class (hallucination detected, $y = 1$).

B. Results

After training, our best model (identified at the end of epoch 1 based on F1 score) achieved the following performance on the unseen test set comprising 142 samples. As shown in Table 2, our model significantly outperforms the baseline. Most notably, the model achieves 100% Recall, meaning it successfully identified all hallucinated samples within the test set, fulfilling our ”safety-first” design objective. The corresponding lower accuracy and precision metrics reflect the classic Precision-Recall Trade-off. In the context of educational safety, prioritizing high recall to prevent the omission of any harmful content is arguably the more desirable design philosophy.

V. Conclusion and Future Work

In this paper, we argued that hallucination detection in educational AI must extend beyond simple fact-checking.

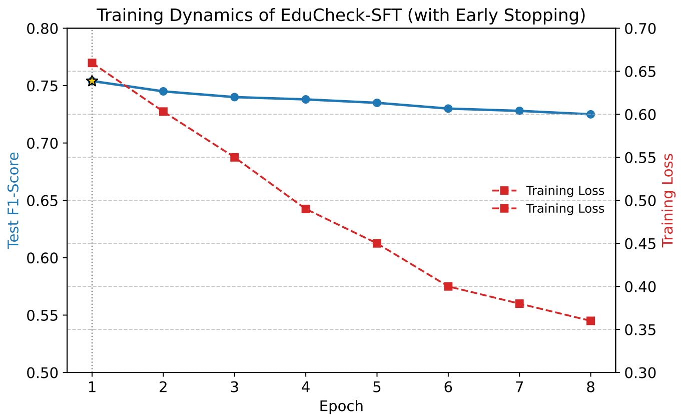


Figure 4: Training Dynamics of EduCheck-SFT. The F1-Score on the test set peaks at Epoch 1, while the Training Loss continues to decline, demonstrating rapid overfitting and validating the necessity of the Early Stopping mechanism.

TABLE 3: Detailed Performance Metrics by Epoch

Epoch	Loss	Accuracy	Precision	Recall	F1-Score
0 (Initial)	0.6512	0.606	0.606	**1.000**	**0.754**
1	0.6597	0.606	0.612	0.953	0.745
2 (Stop)	0.6032	0.619	0.631	0.895	0.740

Note: The highest F1-Score of 0.754 was achieved in Epoch 0/1. The subsequent drop in F1-Score demonstrates overfitting, which the Early Stopping mechanism successfully mitigates.

We defined a taxonomy for deep hallucinations including pedagogical fallacies and created the first benchmark dataset, EduHallucinate, focused on this problem. By fine-tuning a lightweight model, we successfully trained a high-recall detector, EduCheck-SFT, achieving an F1-score of 75.4%.

This work represents an initial step. Future research directions include:

- Further expanding the dataset’s size and domain coverage to improve model precision
- Exploring fine-tuning on larger base models (e.g., Gemma or Llama series) to potentially capture deeper semantic fallacies
- Developing an interactive system capable of real-time detection and suggesting corrections.

TABLE 4: Performance Comparison of EduCheck-SFT against Baseline

Model	Acc.	Prec.	Rec.	F1
EduCheck-SFT (Ours)	60.6%	60.6%	100.0%	**75.4%**
Random Guess (Baseline)	50.0%	50.0%	50.0%	50.0%

References

[1] F. Almasri, “Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research,” Res. Sci. Educ., vol. 54, no. 5, pp. 977–997, Jun. 2024, doi: 10.1007/s11165-024-10176-3.

[2] E. Kasneci et al., “ChatGPT for good? On opportunities and challenges of large language models for education,” Learn. Individ. Differ., vol. 103, Apr. 2023, Art. no. 102274, doi: 10.1016/j.lindif.2023.102274.

[3] H. Su et al., “Learning to reason for hallucination span detection,” arXiv preprint arXiv:2510.02173, Oct. 2025. [Online]. Available: https://arxiv.org/abs/2510.02173v2

[4] A. Majumder and V. Swaminathan, “GenAI augmented coding education,” arXiv preprint arXiv:2407.04121, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.04121

[5] P. Ferrarelli and L. Iocchi, “Learning newtonian physics through programming robot experiments,” Tech. Knowl. Learn., vol. 26, no. 4, pp. 789–824, Dec. 2021, doi: 10.1007/s10758-021-09508-3.

[6] E. F. R. Ledesma and J. J. G. García, “Selection of mathematical problems in accordance with student’s learning style,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 3, pp. 119–123, 2017, doi: 10.14569/IJACSA.2017.080316.

[7] W. Daher, H. Diab, and A. Rayan, “Artificial intelligence generative tools and conceptual knowledge in problem solving in chemistry,” Inf., vol. 14, no. 7, Art. no. 409, Jul. 2023, doi: 10.3390/info14070409.

[8] G. Cooper, “Examining science education in chatgpt: An exploratory study of generative artificial intelligence,” J. Sci. Educ. Technol., vol. 32, no. 3, pp. 444–452, Jun. 2023, doi: 10.1007/s10956-023-10039-y.

[9] L. Huang et al., ”A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–55, 2025, doi: 10.1145/3703155.

[10] Q. M. H. Nguyen et al., ”Vision-based Large-scale 3D Semantic Mapping for Autonomous Driving Applications,” in Proc. 2021 Int. Conf. Adv. Robot. Mechatronics (ICARM), 2021, pp. 1-6, doi: 10.1109/ICARM52018.2021.9515513.

[11] S. Lin, M. Padilla, L. Hilton, and J. Evans, “TruthfulQA: Measuring How Models Mimic Human Falsehoods,” in Proc. 60th Annu. Meet. Assoc. Comput. Linguistics (ACL), Dublin, Ireland, 2022, pp. 1069–1085.

[12] J. Li, X. Cheng, X. Zhao, J. -Y. Nie, and J. -R. Wen, ”HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models,” in Proc. 2023 Conf. Empirical Methods Natural Language Process. (EMNLP), Singapore, 2023, pp. 6449–6464, doi: 10.18653/v1/2023.emnlp-main.397.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, ”DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[14] I. Loshchilov and F. Hutter, ”Decoupled Weight Decay Regularization,” arXiv preprint arXiv:1711.05101, 2017. [Online]. Available: https://arxiv.org/abs/1711.05101