Educational Hallucination Evaluation System

Pranali Kalokhe[1], Mansi Shukla[2], Swapnali Palve[3], Siddhi Patil[4]
[1,2,3,4]Pillai College of Arts Commerce and Science Autonomous

***Abstract*—Generative Artificial Intelligence (AI) is increasingly used in education to answer questions, explain concepts, and support student learning. However, AI-generated responses may sometimes contain incorrect, incomplete, or unsupported information, commonly known as hallucination. Such responses can affect students when incorrect information is accepted as factual.**

This research presents an Educational Hallucination Evaluation System, an evidence-based prototype developed to evaluate educational responses. The system uses 100 educational questions covering different subjects, question types, and difficulty levels. Each question is linked to a reference ground-truth answer and evidence source. Submitted responses are evaluated through text normalization and word-level comparison with the corresponding ground-truth answer.

A Streamlit-based web application was developed to provide an interactive evaluation process. Users can select a question, view its details, enter a response, and receive the similarity percentage, hallucination status, severity level, and reference answer. The research data was analyzed using Python, Pandas, Matplotlib, and statistical methods.

The evaluation classified 88 responses as correct, 8 as partially correct, and 4 as incorrect. A total of 12 responses were classified as hallucinated, resulting in an overall hallucination rate of 12%. Chi-square tests were also conducted to examine the relationship between hallucination and question type, difficulty, and subject.

***Index Terms*—Educational AI, AI Hallucination, Hallucination Detection, Generative AI, Natural Language Processing, Streamlit, Ground Truth, Response Evaluation, Educational Technology.**

I. INTRODUCTION

Artificial Intelligence has become increasingly common in educational applications. Generative AI systems can answer questions, explain concepts, summarize information, and provide learning assistance. This has made AI-based educational tools useful for students and teachers.

However, the ability of generative AI to produce fluent and convincing answers does not guarantee that every generated statement is factually correct. An AI system may generate an answer that appears reasonable but contains incorrect facts, incomplete information, or unsupported claims. This problem is generally known as AI hallucination.

Hallucination is particularly important in education because learners may use AI-generated responses as a source of knowledge. If an incorrect response is accepted without verification, it may lead to misunderstanding of the subject.

Traditional evaluation of AI responses often depends on manual checking. Manual evaluation can require considerable time, especially when a large number of responses need to be analyzed.

Therefore, there is a need for a simple system that can provide a preliminary automated evaluation of educational responses.

The proposed Educational Hallucination Evaluation System addresses this requirement by using a reference-based evaluation approach. The system maintains a ground-truth dataset containing questions, correct answers, subject information, difficulty levels, and evidence sources. A submitted response is compared with the corresponding reference answer and classified according to predefined similarity thresholds.

The research follows a structured approach involving dataset preparation, data quality checking, response evaluation, statistical analysis, and deployment through a Streamlit web interface.

II. PROBLEM STATEMENT

Generative AI can produce educational answers that are grammatically correct and convincing but may not always be factually accurate. Identifying such incorrect responses manually can be time-consuming, particularly when many educational questions and AI-generated responses are involved.

The main problem addressed by this research is:
How can an evidence-based evaluation system be developed to provide a simple and consistent assessment of educational responses and identify potential hallucinations using reference ground-truth information?
The system focuses on comparing responses with reference answers and providing measurable evaluation results rather than relying only on subjective manual judgment.

III. RESEARCH OBJECTIVES

The main objectives of this research are:
- To develop a reference-based system for evaluating educational responses.
- To create and organize a dataset containing 100 educational questions.
- To associate each question with a ground-truth answer and evidence source.
- To evaluate responses using text normalization and word-level similarity.
- To classify responses into Correct, Partially Correct, and Incorrect categories.

IV. LITERATURE REVIEW

AI-based educational systems support personalized learning and automated assistance, but the reliability of AI-generated information remains a concern. AI responses may appear convincing while containing incorrect or unsupported information, so fluency alone cannot ensure factual accuracy.
A reference-based evaluation approach addresses this issue by comparing an AI response with a verified ground-truth answer. The proposed system follows this approach and evaluates responses using similarity, correctness, hallucination status, severity, and category-wise analysis.
The study presents a prototype-level evaluation method implemented using Python data-analysis tools and the Streamlit framework.

V. PROPOSED METHODOLOGY

5.1. Requirement Analysis

The requirement analysis identifies the main needs of the Educational Hallucination Evaluation System.
The system requirements are:
- Educational Question Dataset – A dataset containing 100 questions with question type, subject, and difficulty.
- Ground-Truth Answers – A verified reference answer and evidence source for each question.
- Response Evaluation – Ability to compare a submitted response with the corresponding ground-truth answer.
- Similarity Calculation – Calculate the percentage of matching information between the response and ground truth.
- Hallucination Detection – Identify whether the response is potentially hallucinated based on predefined similarity thresholds.

5.2. System Design

The system is designed as a modular, reference-based evaluation system. Its main components are:
1. Question Selection Module – Allows the user to select a question ID.
2. Question Information Module – Displays the selected question, question type, subject, and difficulty.
3. Response Input Module – Allows the user to enter an educational response.
4. Text Processing Module – Converts the response and ground-truth answer into a standardized form for comparison.
5. Similarity Evaluation Module – Compares the response with the ground-truth answer and calculates similarity.
6. Classification Module – Classifies the response as Correct, Partially Correct, or Incorrect.
7. Hallucination and Severity Module – Determines hallucination status and assigns severity.
8. Result Display Module – Displays similarity, evaluation, hallucination, severity, ground truth, and evidence source.

The system was implemented using Python and Streamlit. The datasets were processed and analyzed using Python-based data analysis libraries.

5.3. Implementation

The implementation includes:
1. Loading the educational question and ground-truth datasets.

2. Selecting questions through the Streamlit interface.
3. Displaying question details such as type, subject, and difficulty.
4. Accepting the user's response through a text input area.
5. Cleaning and normalizing the response and ground-truth text.
6. Comparing the words in both texts to calculate similarity.
7. Applying predefined thresholds to classify the response.
8. Identifying hallucination and assigning severity.
9. Displaying the evaluation result and reference ground truth.
10. Performing further analysis using the evaluation dataset.
11. Generating research results in CSV and Excel formats.

The project uses files such as educational_hallucination_100_questions.csv, ground_truth_completed.csv, AI_response_evaluation_filled.csv, andhallucination_analysis_ results.xlsx.

5.4. Testing and Evaluation

Testing was conducted to ensure that the system works correctly and produces consistent results. The evaluation included dataset validation, Question ID checking, response comparison, similarity calculation, classification, hallucination detection, severity assignment, statistical testing, and Streamlit interface testing.

The system was tested using 100 responses. The results showed 88 Correct, 8 Partially Correct, and 4 Incorrect responses. Based on the defined evaluation rules, 12 responses were classified as hallucinated. These results were further analyzed using chi-square tests for question type, difficulty, and subject.

VI. SYSTEM ARCHITECTURE

6.1. User Input Layer

The User Input Layer allows the user to select a question and enter an answer in the Streamlit interface. The selected question displays its Question ID, question, question type, subject, and difficulty level. The user then provides a response, which is passed to the evaluation module for further processing.

6.2. Rule-Based Validation Module

The Rule-Based Validation Module evaluates the submitted response by comparing it with the corresponding ground-truth answer. The system first normalizes the text by converting it to lowercase, removing punctuation and unnecessary characters, and separating the words. The processed response is then compared with the reference answer to determine the level of matching.

6.3. Similarity Calculator

The Similarity Calculator measures the overlap between the submitted response and the ground-truth answer. The system identifies common words and calculates a similarity percentage based on the number of matching words in relation to the ground-truth answer. This value is used as one of the main inputs for response classification.

6.4. Response Classification Module

The Response Classification Module applies predefined thresholds to the calculated similarity. Responses with 80% or above similarity are classified as Correct, responses between 35% and 79.99% are classified as Partially Correct, and responses below 35% are classified as Incorrect.

6.5. Hallucination and Severity Module

This module identifies whether the response contains a potential hallucination based on the evaluation result. Correct responses are marked as No Hallucination, while partially correct and incorrect responses are classified as hallucinated according to the defined rules. Severity is assigned as None, Minor, or Major depending on the evaluation category.

6.6. Feedback Module

The Feedback Module provides the evaluation result to the user in an understandable format. It displays the similarity percentage, evaluation category, hallucination status, severity level, and reference/ground-truth answer. This helps the user understand how closely their response matches the expected educational answer.

6.7. Real-Time Display Module

The Real-Time Display Module presents the evaluation results directly through the Streamlit interface after the user submits a response. The results are displayed

immediately without
requiring a separate application. This makes the prototype simple and interactive for testing educational responses.

VII. EVALUATION STRATEGY

7.1. Accuracy Testing

Accuracy testing was performed to check whether the system correctly evaluates educational responses against the corresponding ground-truth answers. The system was tested using 100 responses and classified them into Correct, Partially Correct, and Incorrect categories based on the predefined similarity thresholds.

The results showed 88 Correct responses (88%), 8 Partially Correct responses (8%), and 4 Incorrect responses (4%). The system also identified 12 responses as potential hallucinations (12%). These results indicate that the reference-based evaluation method was able to provide consistent classifications for the tested responses.

7.2. Performance Testing

Performance testing was conducted to verify the responsiveness and usability of the Streamlit-based system. The system loads the question and ground-truth datasets, accepts the user's response, performs text preprocessing and similarity calculation, and displays the evaluation results.

The evaluation process is lightweight because it uses text normalization, word-level comparison, and predefined rules rather than a computationally intensive machine-learning model. The results are displayed directly in the Streamlit interface after submission, making the prototype suitable for interactive educational response evaluation.

VIII. RESULTS AND DISCUSSION

8.1. Overall Results

The system was tested using 100 educational responses. The results showed 88 Correct (88%), 8 Partially Correct (8%), and 4 Incorrect (4%) responses. Overall, the system successfully classified the responses based on their similarity with the ground-truth answers.

8.2. Hallucination Results

Out of 100 responses, 88 responses (88%) were classified as having no hallucination, while 12 responses (12%) were identified as potential hallucinations. This shows how the reference-based approach can be used to identify responses that differ from the expected ground-truth information.

8.3. Severity Results

The severity analysis showed 88 responses with None severity (88%), 8 with Minor severity (8%), and 4 with Major severity (4%). This classification helps distinguish between responses that match the reference answer and those that show partial or incorrect information.

8.4. Statistical Discussion

The relationship between hallucination and question type, difficulty, and subject was tested using the chi-square test. The obtained p-values were 0.824, 0.137, and 0.474, respectively. Since all values are greater than 0.05, no statistically significant association was found in the tested dataset.

8.5. Discussion

The results show that the proposed system provides a structured method for evaluating educational responses using ground-truth answers, similarity calculation, rule-based classification, hallucination detection, and severity analysis. However, the system depends on word-level comparison and the quality of the reference answers. Therefore, the results should be considered as a prototype evaluation rather than a complete semantic fact-checking system.

IX. KEY CONCEPTUAL OBSERVATIONS

1. Importance of Ground Truth:
Ground-truth answers provide a reliable reference for comparing and evaluating educational responses.

2. Reference-Based Evaluation:
The system uses text similarity and predefined rules to classify responses as Correct, Partially Correct, or Incorrect.

3. Hallucination Identification:
Responses with insufficient similarity to the ground-truth answer can be identified as potential hallucinations.

4. Question Characteristics:
Hallucination percentages vary across different question types, difficulty levels, and subjects in the tested dataset.

5. Statistical Findings:
The chi-square analysis did not show a statistically significant association between hallucination and question type, difficulty, or subject.

6. Need for Advanced Methods:
The current word-level comparison has limitations in understanding meaning and context. Advanced NLP techniques could improve semantic evaluation.

7. Practical Application:
The Streamlit interface provides a simple and interactive way to evaluate educational responses and display similarity, evaluation, hallucination status, severity, and reference answers.

X. CONCLUSION

The Educational Hallucination Evaluation System provides a simple reference-based approach for evaluating educational responses and identifying potential hallucinations. The system compares user responses with verified ground-truth answers using text preprocessing, similarity calculation, and predefined classification rules.

The evaluation of 100 responses resulted in 88% Correct, 8% Partially Correct, and 4% Incorrect responses, with 12% identified as potential hallucinations. The system also provides severity classification and an interactive Streamlit interface for displaying the results.

Overall, the project demonstrates that a ground-truth-based approach can provide a structured and practical way to evaluate educational responses. However, the current word-level comparison has limitations in understanding deeper meaning and context. Future improvements can include advanced NLP and semantic techniques to make the evaluation more accurate and comprehensive.

XI. FUTURE WORK

- Advanced NLP Techniques:

Future versions can use semantic similarity, embeddings, and transformer-based models to understand the meaning of responses more accurately.

- Improved Hallucination Detection:

The system can be enhanced to identify specific incorrect or unsupported claims instead of relying mainly on word-level similarity.

- Larger Dataset:

More questions from different subjects, difficulty levels, and question types can be added to improve the reliability of the evaluation.

- Multiple Reference Sources:

Multiple verified sources can be used instead of a single ground-truth answer to provide more comprehensive evaluation.

- Detailed Feedback:

The system can provide explanations about which parts of a response are correct, incomplete, or potentially hallucinated.

- Improved Streamlit Dashboard:

Future development can include interactive charts, detailed statistics, response history, and downloadable evaluation reports.

- Automated Evaluation:

Advanced AI-based evaluation methods can be integrated to improve semantic understanding and provide more detailed educational feedback.

XII. PLAGIARISM



REFERENCES

[1] Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Computing Surveys, 2023.

[2] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[3] P. Lewis et al., “Retrieval-augmented

generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[4] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[5] W. McKinney, Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

[6] pandas Development Team, pandas Documentation. Python Data Analysis Library. [Online]. Available: https://pandas.pydata.org/docs/.

[7] Streamlit, Inc., Streamlit Documentation. [Online]. Available: https://docs.streamlit.io/.

[8] SciPy Community, SciPy Documentation. Scientific Python Ecosystem. [Online]. Available: https://docs.scipy.org/doc/scipy/.

[9] "PassGuard: Design and Evaluation of a Real-Time Password Strength Checker," research paper used as a reference for research structure and methodology organization. Publication details not provided.