584 Final Project Report

Title: Evaluating Automated Systems for Detecting Faulty Science Questions: Classifier Performance, Disciplinary Differences, and Faulty Reason Types

1)Introduction

Large Language Models (LLMs) have demonstrated extraordinary capabilities in generating human-like responses and addressing complex queries across a wide array of disciplines. Despite their remarkable achievements, a significant vulnerability persists: the tendency of LLMs to hallucinate. Hallucination refers to situations where LLMs generate outputs that are factually incorrect, logically inconsistent, or entirely fabricated.(Huang et al., 2024) Moreover, LLMs may try to calculate or answer even if the question itself is faulty. This issue becomes particularly concerning in domains requiring precision, such as science, where erroneous outputs can mislead users and compromise trust in these systems.

The root causes of hallucination in LLMs lie in their training paradigms. LLMs are trained on vast datasets that may contain errors, ambiguities, or incomplete information. Furthermore, these models lack an inherent understanding of the world, relying instead on statistical patterns, which can lead to overconfidence in generating answers to faulty or misleading questions. This makes LLMs particularly susceptible to queries with inherent contradictions, logical fallacies, or subtle inaccuracies.

To address this weakness, recent works have explored various methods. For instance, DecoPrompt leverages LLMs to "decode" the false-premise prompts without really eliciting hallucination output from LLMs. (Xu, 2024) In addition, pre-training can help with the problem by collecting and emphasizing credible text sources in the pre-training corpus. (Touvron, 2023) While these approaches have shown promise, they often focus on improving response quality rather than identifying faulty or misleading questions at the beginning.

This project presents an automated system designed to detect faulty science questions before they are processed by LLMs. There are 3 main research questions this project is exploring: (1)What machine learning classifier trained with labeled data performs the best to identify faulty questions? (2)Do classifiers perform differently on faulty questions in conceptual disciplines (e.g., biology) versus computational disciplines (e.g., mechanical engineering)? (3)Do faulty reason types affect classifiers' performance?

2)Dataset Curation

The dataset for this research project consists of faulty science questions curated across 20 distinct subjects, with each subject contributing 5 original questions and 5 similar variations, resulting in

a total of 200 unique questions. These questions are specifically designed to challenge large language models (LLMs) , specifically ChatGPT by embedding subtle errors or logical inconsistencies. The dataset includes questions from diverse disciplines such as biology and mechanical engineering, ensuring a broad spectrum of challenges for automated detection systems.

The original correct science questions were sourced from two publicly available datasets:

1. **StemQ:** A dataset for learning university STEM courses at scale and generating questions at a human level. (Drori et al., 2023)
2. **ScienceQA:** A new benchmark that consists of 21,208 multimodal multiple choice questions with a diverse set of *science* topics and annotations of their answers with corresponding *lectures* and *explanations*. (Lu et al., 2022)

These original questions were meticulously edited to introduce faults and test the susceptibility of LLMs to hallucination. The faults included definition fallacies, value contradictions, and misrepresentations designed to resemble plausible scientific queries.

This project mainly used the following columns in the dataset:

Discipline: The subject or field of the question (e.g., Biology, Mechanical Engineering).

Question: The edited faulty science question.

Original correct question: The original question before being edited to introduce faults.

Faulty Reason Type: The type of fault introduced.

The dataset was constructed in 2 stages:

1. Original Dataset (data100.csv):
    - This dataset contains 100 rows of data, where the 20 subjects were merged and shuffled.
    - Each question has a corresponding "Original correct question" entry, ensuring clear traceability of edits.
2. Augmented Dataset (data200.csv):
    - To expand the dataset, each question was further augmented by creating variations with slight modifications in wording or context, resulting in a total of 200 rows.
    - The augmented dataset merges all subjects and shuffles the questions to prevent any subject-specific bias during evaluation.

This curated dataset forms the foundation for evaluating the automated detection system, providing a rigorous and diverse set of faulty queries to challenge and improve the reliability of LLMs.

3)Research Questions and Experiments

Evaluation Metrics:

This project employed multiple evaluation metrics to comprehensively assess the performance of classifiers in identifying faulty science questions. **Accuracy** served as the primary metric, representing the proportion of correctly classified questions among all samples. However, relying solely on accuracy can be misleading in the presence of class imbalance. To address this, additional metrics were used: **precision**, which quantifies the proportion of correctly identified faulty questions among all predictions labeled as faulty; **recall**, which measures the ability to detect all actual faulty questions; and **F1-score**, the harmonic mean of precision and recall, providing a balanced view of a model's performance. These metrics together ensure a thorough evaluation of the models, particularly in scenarios where the cost of false positives or false negatives may vary. By combining these metrics, the project provides nuanced insights into the strengths and weaknesses of both traditional and advanced classifiers.

3.1 Research Question 1

Overview:

This subsection is exploring the first research question: “What machine learning classifier trained with labeled data performs the best to identify faulty questions?” The goal of this experiment is to evaluate the performance of traditional and advanced machine learning classifiers in identifying faulty science questions. Specifically, the experiment compares the accuracy, precision, recall, and F1-score of traditional classifiers (Logistic Regression, Random Forest, and Support Vector Machine) against advanced models (BERT, RoBERTa, RNN, and LSTM) trained on labeled data.

Dataset and environment:

The original dataset containing 200 labeled examples of faulty and correct science questions was used to create a new dataset with 2 columns. The "Question" column served as the input text. The "Faulty" column, where 1 indicates a faulty question and 0 indicates a correct question, was used as the target. The dataset was split into 80% training and 20% testing data using stratified sampling to maintain class balance. All models were trained on Google Colab environment.

Models Evaluated:

-Traditional Classifiers

Logistic Regression:

- A baseline linear model suitable for binary classification.
- TF-IDF (Term Frequency-Inverse Document Frequency) vectors were used as input features.

Random Forest:

- An ensemble model that uses multiple decision trees for classification.
- Input: TF-IDF vectors.

Support Vector Machine (SVM):

- A robust linear classifier optimized for separating hyperplanes.
- Input: TF-IDF vectors.

-Advanced Models

BERT (Bidirectional Encoder Representations from Transformers):

- Pretrained transformer model fine-tuned on the labeled dataset.
- Tokenized inputs using bert-base-uncased.

RoBERTa (Robustly Optimized BERT Pretraining Approach):

- An optimized version of BERT fine-tuned for binary classification.
- Tokenized inputs using roberta-base.

RNN (Recurrent Neural Network):

- A simple recurrent neural network model trained with word embeddings.

LSTM (Long Short-Term Memory):

- An advanced RNN variant designed to handle long-term dependencies, trained with word embeddings.

Experimental Setup:

1. Preprocessing:
    - For traditional models: Questions were tokenized and transformed into TF-IDF vectors with a vocabulary size of 10,000.
    - For advanced models: Tokenization and input preparation followed model-specific pipelines (e.g., Hugging Face Transformers for BERT and RoBERTa, and Keras for RNN/LSTM).
2. Training Configuration:
    - Traditional Models: Grid search was used to optimize hyperparameters (e.g., regularization strength for Logistic Regression, number of estimators for Random Forest, and kernel type for SVM).
    - Advanced Models: Fine-tuned pretrained models (BERT, RoBERTa) for three epochs with a batch size of 16 and a learning rate of 2e-5. RNN and LSTM models were trained for five epochs with a learning rate of 1e-3.

Results:

Model	Accuracy	precision	recall	f1-score
Logistic Regression	0.23	0.21	0.23	0.22
Random Forest	0.53	0.54	0.54	0.52
Support Vector Machine	0.26	0.24	0.28	0.24
BERT	0.69	0.76	0.67	0.65
RoBERTa	0.73	0.81	0.74	0.72
RNN	0.46	0.46	0.46	0.45
LSTM	0.47	0.24	0.47	0.32

-Traditional Classifiers:

- Logistic Regression performed poorly across all metrics, achieving an accuracy of 0.23, with precision, recall, and F1-score hovering around 0.22. This indicates that the model struggled to distinguish between faulty and correct questions effectively.
- Random Forest showed moderate performance with an accuracy of 0.53 and balanced precision and recall at 0.54, resulting in an F1-score of 0.52. This suggests Random Forest could capture some patterns but lacked consistency across the dataset.
- Support Vector Machine (SVM) also performed poorly, with an accuracy of 0.26 and an F1-score of 0.24. While it slightly outperformed Logistic Regression in recall (0.28), it remained ineffective overall.

-Advanced Models:

- BERT significantly outperformed traditional models, achieving an accuracy of 0.69. It demonstrated strong precision (0.76) and recall (0.67), resulting in an F1-score of 0.65. This highlights its ability to leverage contextual information effectively for classification tasks.
- RoBERTa was the best-performing model, with an accuracy of 0.73 and the highest precision (0.81), recall (0.74), and F1-score (0.72). This demonstrates the advantage of its optimized architecture in handling nuanced and contextually complex queries.
- RNN delivered mediocre results, with an accuracy of 0.46 and nearly identical precision, recall, and F1-score around 0.45, indicating limited capability to capture sufficient patterns for this task.
- LSTM slightly improved upon RNN in accuracy (0.47) but had uneven metric scores, including a low precision of 0.24 and a recall of 0.47. Its F1-score of 0.32 suggests it struggled to balance false positives and negatives.

3.2 Research Question 2

Overview:

The goal of this experiment is to determine whether machine learning classifiers perform differently on faulty questions originating from conceptual disciplines (e.g., biology) compared to computational disciplines (e.g., mechanical engineering). The experiment evaluates whether the nature of the discipline influences the classifier's ability to detect faulty questions.

Dataset and environment:

20 disciplines were categorized into 2 fields:

Conceptual Disciplines: Biology, Chemistry, Earth Science, Physics, Social Science, Astronomy, and Language Science.

Computational Disciplines: Mechanical Engineering, Mathematics, Computer Science, Electrical Engineering, Statistics, Quantum Physics, and Units & Measurement.

The dataset used in experiment 1 was split into two subsets based on the "Discipline" column: conceptual and computational disciplines. Each subset was further divided into training (80%) and testing (20%) sets. All models were trained on Google Colab environment.

Models Evaluated:

The best performing traditional and advanced classifiers, which are Random Forest and RoBERTa, were used here.

Experimental Setup:

Models were trained and evaluated separately on the conceptual and computational subsets. For the transformer-based model (RoBERTa), fine-tuning was performed using discipline-specific subsets. Traditional model was trained using TF-IDF features.

Results:

Model	Discipline Type	Accuracy	precision	recall	f1-score
Random Forest	Conceptual	0.57	0.61	0.59	0.57
Random Forest	Computational	0.47	0.51	0.51	0.43
RoBERTa	Conceptual	0.53	0.49	0.49	0.45
RoBERTa	Computational	0.78	0.85	0.75	0.75

Random Forest:

- For conceptual disciplines, Random Forest achieved an accuracy of 0.57, with balanced precision (0.61) and recall (0.59), resulting in an F1-score of 0.57. This suggests that Random Forest effectively captures patterns in conceptual questions but has limited precision in identifying faulty queries.
- In computational disciplines, the model performed worse, with an accuracy of 0.47 and an F1-score of 0.43. Precision and recall were both 0.51, indicating difficulty in handling computational questions compared to conceptual ones.

RoBERTa:

- For conceptual disciplines, RoBERTa demonstrated moderate performance, with an accuracy of 0.53 and an F1-score of 0.45. Precision and recall were nearly equal at 0.49, reflecting its ability to identify faulty conceptual questions, albeit with limited effectiveness.
- In computational disciplines, RoBERTa significantly outperformed both Random Forest and its performance in conceptual disciplines, achieving an accuracy of 0.78 and an F1-score of 0.75. Precision was particularly high at 0.85, indicating that RoBERTa excelled in correctly identifying faulty computational questions, though recall was slightly lower at 0.75.

3.3 Research Question 3

Overview:

The purpose of this experiment is to evaluate whether the type of fault introduced in a science question (e.g., faulty definition) affects the performance of machine learning classifiers. By focusing on the two most common faulty reason types, this experiment aims to determine whether classifiers are equally effective at identifying different types of faults or if certain types pose unique challenges.

Dataset and environment:

Two most common faulty reason types were chosen:
Faulty Definition: Questions with incorrect or misleading definitions.
Faulty Value: Questions containing numerical or quantitative errors.

The dataset was filtered to include only rows labeled with the "Faulty Reason Type" column values of faulty definition or faulty value. The filtered dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to preserve class distributions. All models were trained on Google Colab environment.

Models Evaluated:
Similar to experiment 2, the best performing traditional and advanced classifiers, which are Random Forest and RoBERTa, were used here.

Experimental Setup:

Each classifier was trained and evaluated separately on subsets corresponding to each faulty reason type (faulty definition and faulty value). For transformer-based models (RoBERTa), fine-tuning was performed using the respective subsets for three epochs with a learning rate of 2e-5. For traditional models, TF-IDF features were used as input.

Results:

Model	Faulty Reason Type	Accuracy	precision	recall	f1-score
Random Forest	Faulty Definition	0.51	0.55	0.54	0.51
Random Forest	Faulty Value	0.42	0.41	0.42	0.4
RoBERTa	Faulty Definition	0.42	0.21	0.3	0.18
RoBERTa	Faulty Value	0.5	0.25	0.5	0.33

Random Forest:

- For faulty definition, Random Forest achieved an accuracy of 0.51 with balanced precision (0.55) and recall (0.54), resulting in an F1-score of 0.51. This indicates that Random Forest was moderately effective at detecting faulty definitions, leveraging linguistic patterns inherent in these faults.
- For faulty value, the model's performance declined, with an accuracy of 0.42 and an F1-score of 0.40. Precision and recall were closely matched at 0.41 and 0.42, respectively, suggesting that Random Forest struggled more with numerical inconsistencies compared to definitions.

RoBERTa:

- For faulty definition, RoBERTa showed limited effectiveness, with an accuracy of 0.42 and an F1-score of 0.18. Precision (0.21) and recall (0.30) were notably low, indicating difficulty in identifying faulty definitions despite its advanced contextual understanding.
- For faulty value, RoBERTa slightly outperformed its performance on faulty definitions, achieving an accuracy of 0.50 and an F1-score of 0.33. Recall was relatively high at 0.50, but precision remained low at 0.25, indicating that the model was prone to false positives in this category.

4)Discussion

Q1:

The results indicate that advanced transformer-based models (BERT and RoBERTa) significantly outperformed traditional classifiers and RNN-based architectures. Among traditional methods, Random Forest provided the highest accuracy and balanced metrics, although it still fell short compared to advanced models. The superior performance of RoBERTa, with its optimized architecture and training techniques, underscores the importance of leveraging pretrained language models for complex classification tasks involving subtle contextual nuances. Meanwhile, RNN and LSTM models, despite their ability to process sequential data, showed limited effectiveness, likely due to the small dataset size and lack of deeper contextual understanding compared to transformers.

These findings highlight the critical role of advanced architectures in identifying faulty questions and suggest that future efforts should prioritize transformer-based approaches for enhanced reliability and robustness.

Q2:

Discipline Type Impact:

Both models exhibited better performance in conceptual disciplines in terms of F1-score, though RoBERTa's accuracy in computational disciplines was much higher (0.78 compared to 0.53 for conceptual). This suggests that computational questions are generally easier for RoBERTa to classify, potentially due to their structured nature and the pretrained model's strength in handling such contexts.

Conversely, Random Forest showed stronger performance in conceptual disciplines, likely due to its reliance on feature patterns that align well with the more descriptive nature of conceptual questions.

Model Comparison:

While Random Forest demonstrated balanced performance across conceptual and computational disciplines, its overall accuracy and F1-scores were lower than RoBERTa's performance in computational disciplines.

RoBERTa significantly outperformed Random Forest in computational disciplines, showcasing the advantage of transformer-based models in leveraging context and structure effectively.

Q3:

Faulty Reason Type Impact:

Faulty definitions were better handled by Random Forest, which showed higher accuracy, precision, recall, and F1-scores compared to RoBERTa. This suggests that traditional models like Random Forest may be better suited for tasks involving simpler, pattern-based linguistic inconsistencies.

Faulty values, while more challenging overall, saw comparable accuracy between Random Forest and RoBERTa. RoBERTa's higher recall indicates its ability to detect faulty value cases, albeit at the cost of lower precision.

Classifier Differences:

Random Forest outperformed RoBERTa on faulty definitions, likely due to its reliance on decision-tree patterns that align better with the nature of definition-related faults.

RoBERTa, despite being a transformer-based model, struggled with faulty definitions. Its higher recall on faulty values suggests potential strength in identifying numerical inconsistencies but indicates room for improvement in precision and general performance.

Small Dataset Size:

The limited dataset size (less than 200 samples for each type) likely contributed to the variability in results, particularly for RoBERTa. Transformer-based models typically require larger datasets to generalize effectively, which may explain their relatively poor performance in this experiment.

5)Conclusion and Future Work

This study evaluated automated systems for detecting faulty science questions, focusing on classifier performance across different dimensions: overall effectiveness, disciplinary differences, and fault type variations. From the experiments, it's easy to conclude that advanced models like RoBERTa outperformed traditional models, particularly in tasks requiring nuanced contextual understanding. Random Forest, while moderately effective, was better suited for simpler, pattern-based tasks such as detecting faulty definitions. Moreover, conceptual disciplines posed more challenges for classifiers, with RoBERTa performing significantly better on computational disciplines due to their structured nature. Traditional models like Random Forest were more balanced in handling conceptual queries but lacked the depth to compete with transformer-based models. Finally, faulty definitions were more effectively handled by traditional classifiers like Random Forest, which could capture linguistic patterns. Faulty values proved more challenging for all models, but RoBERTa demonstrated better recall, albeit at the cost of lower precision. These findings underscore the importance of selecting classifiers based on the specific characteristics of the task and dataset.

For future directions, expanding the dataset to include more samples and diverse faulty question types will improve the reliability and generalizability of the results. Also, exploring hybrid models that combine traditional decision trees with transformer-based embeddings may address weaknesses in both approaches.

References

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Office Information Systems*. https://doi.org/10.1145/3703155

Xu, N., & Ma, X. (2024). *DecoPrompt: Decoding prompts reduces hallucinations when large language models meet false premises*. arXiv. https://arxiv.org/abs/2411.07457

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv. https://arxiv.org/abs/2307.09288

Drori, I., Zhang, S., Chin, Z., Shuttleworth, R., Lu, A., Chen, L., Birbo, B., He, M., Lantigua, P., Tran, S., Hunter, G., Feng, B., Cheng, N., Wang, R., Hicke, Y., Surbehera, S., Raghavan, A., Siemenn, A., Singh, N., … Solar-Lezama, A. (2023). A dataset for learning university STEM courses at scale and generating questions at a human level. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(13), 15921–15929. https://doi.org/10.1609/aaai.v37i13.27091

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., & Kalyan, A. (2022). *Learn to explain: Multimodal reasoning via thought chains for science question answering*. arXiv. https://arxiv.org/abs/2209.09513