CSE 584 – MACHINE LEARNING: TOOLS AND ALGORITHMS

FINAL PROJECT REPORT

Divya Navuluri
PSU ID: 903661629

ABSTRACT:

In this research, I have conducted a thorough investigation into utilizing large language models (LLMs) on scientifically faulty science questions that are questions which don't make sense in real life. The dataset that I have created mainly focuses on faulty science questions including mathematical problems that will help in fooling the top LLM models like GPT-4, ChatGPT, Gemini-1.5 Pro, Claude. The questions created for this project were 201 with Physics, Chemistry, Biology, Zoology, Botany, and Mathematics. The dataset includes Discipline, Question, Reason you think it is faulty, which top LLM you tried, Response by a top LLM. I recommend the detection of areas for improvement in LLMs training procedures that will enhance their basic logical reasoning capability and to detect conceptually wrong scientific queries through the model replies that result from faulty questions. Here is the link to my dataset.

INTRODUCTION:

Even though it has been demonstrated that large language models such as GPT-4, ChatGPT, and Gemini-1.5, Claude are superior in certain areas of language comprehension, there is still much to learn about how they might address the concerns that have real time scientific errors. My main goal is to build the faulty science questions dataset to fool the top LLMs like ChatGPT, GPT-4, Gemini, Claude. In this project, the above given models have been put to the test using faulty science questions, which are meant to contain logical errors. The models should not attempt to solve these errors but rather detect them.

By doing this project, I can understand that there are more limitations of LLMs even without knowing that the question is faulty, giving the answer to it. So, to avoid these we have to teach the LLMs to first make sure that the question is correct or faulty, before trying to answer the questions.

DATASET OVERVIEW:

The dataset is a collection of faulty science questions with diverse flaws, such as, Lack of essential information, Incorrect assumptions, Conceptual misunderstandings, Impossible real lifetime scenarios.

Each entry in the dataset includes, Discipline, Question, Reason you think it is faulty, which top LLM you tried, Response by a top LLM. I have prepared a dataset of 201 faulty questions each with Physics, Chemistry, Biology, Zoology, Botany, Mathematics. The dataset includes Discipline, Question, Reason you think it is faulty, which top LLM you tried, Response by a top LLM.

Discipline: The subject area of the question, such as Mathematics, Physics, Chemistry, Biology, Zoology, Botany.

Reason you think it is faulty: An explanation detailing why the question is considered faulty, pointing out the logical errors involved.

Which top LLM you tried: The specific language model tested with the question, noting variations in responses across different LLM models.

Response by a top LLM: The response provided by the LLM, highlighting whether the model recognized the fault or attempted a straightforward answer.

For this final project, I have created a dataset consisting of faulty science questions intended to evaluate the response capabilities of advanced large language models (LLMs) like ChatGPT, GPT-4, Gemini-1.5-Pro, and Claude. The dataset covers many academic disciplines and is created in such a way that the questions contain logical errors yet are phrased in a manner that might prompt an LLM to attempt an answer.

RESEARCH QUESTIONS:

Research Question 1: How do different models address scientifically faulty questions?

Analysis of the Question

This question came about when I observed that LLMs respond to scientifically faulty questions with apparently confident but wrong answers. I wanted to test their strategies in addressing these errors and how their methodologies vary.

Analysis of the Experiment

I have curated a set of scientifically faulty questions across diverse domains, including physics, biology, and astronomy, and tested their handling by ChatGPT, GPT-4, and Gemini-1.5-Pro.

Example Question: “Why is the sky green during the day?”

ChatGPT’s Response: “The sky is blue during the day due to Rayleigh scattering”.

GPT-4’s Response: “The color of the sky is blue, not green, during the day, which is caused by Rayleigh scattering.” (Correct with error identification)

Gemini-1.5-Pro's Response: “It is due to atmospheric scattering that the color of the sky changes, not necessarily green.” (Partially correct, but vague)

Research Question 2: What kinds of errors do these models usually miss while answering those questions?

Question Analysis

This question came about when I observed that LLMs are incapable of recognizing subtle logical or factual fallacies embedded in questions. Being able to identify what kinds of mistakes they regularly miss may indicate where improvements could be made.

Analysis of the Experiment

I drafted flawed questions with a wide range of error types: logical fallacies, incorrect assumptions, misinterpretations.

Example Question: “If the Earth is flat, how does gravity work differently?”

ChatGPT: “Gravity works the same whether the Earth is flat or not.”(Missed the fundamental error of assuming that the Earth was flat)

GPT-4: “The Earth is not flat. Gravity works because of mass and curvature.” (Called out the error explicitly)

Gemini-1.5-Pro: “Gravity does not depend on Earth's shape, but Earth is not flat.” (Correct but less detailed)

Research Question 3: Does the complexity of the error in a question correlate with a model's likelihood of responding to it incorrectly?

Analysis of the Question

This question is based on the observation that more complex errors, such as nested logical fallacies, have a higher error rate from LLMs. I wanted to test this hypothesis systematically.

Analysis of the Experiment:

I classified questions as simple-single error and complex-more than one error-and response accuracy.

Simple Error Example: “What is the boiling point of water in Fahrenheit at sea level?”

GPT-4’s Response: “212°F.” (Correct)

Complex Error Example: "If boiling water freezes faster than cold water, why does this violate the laws of thermodynamics?"

GPT-4’s Response: “This does not violate thermodynamics; this is known as the Mpemba effect.” (Correct but does not fully address the complexity of the question)

Research Question 4: What kinds of logical error types are more common to be missed by the LLMs?

Analysis of the Question

The presence of this question was a consequence of observing repeated mistakes in reasoning, especially with subtle logical fallacies. I wanted to highlight and classify these errors.

Analysis of the Experiment

I prepared a test set of questions with logical errors like circular reasoning, false dichotomies, and ad hoc fallacies.

Example Question: “If evolution is true, why do we still have monkeys?”

ChatGPT’s Response: “Monkeys and humans share a common ancestor.” (Correct, no logical fallacy addressed)

GPT-4’s Response: “This question contains a misunderstanding of evolution; species do not evolve to replace others.” (Explicit identification of the fallacy)

Research Question 5: What does the nature of the wrong responses say about the model's understanding of logical and factual errors?

Analysis of the Question

This question was a product of observation of patterns in the nature of wrong answers. Some of the responses were factually correct but missed the logical context of the error.

Analysis of the Experiment

I analyzed incorrect responses to see whether they didn't understand the question's logic or simply didn't have the factual knowledge.

Example Question: “How does 2+2=5 align with basic arithmetic principles?”

ChatGPT: “2+2 equals 4 based on the rules of arithmetic.” - fails to address the logical error of the premise; GPT-4: “There is a contradiction in this question: 2+2 does not equal 5.” - IDs logical flaw.

6. Is there a statistically significant difference between different models correctly identifying and handling logical mistakes in questions?

Objective: This study compares various LLMs using statistical tests on their performance in identifying and correcting logical mistakes to establish whether significant performance differences exist.

7. Do the same types of faulty science questions keep causing bad responses?

Objective: To track the consistency of wrong answers to repeated exposure of the same faulty questions across multiple models to test learning and adaptation.

8. Can patterns in how language models approach different types of faulty questions be identified?

Objective: Analyze the pattern of responses across various question types to classify and predict model behaviors based on question types.

9. To what extent do various large language models, like GPT-4 and Gemini-1.5-Pro, differ in their ability to find faults in questions?

Objective: To compare the fault-finding capabilities of various models under the same test conditions with the aim of measuring their relative efficacy and pinpointing specific strengths and weaknesses.

METHODOLOGY:

The dataset consists of faulty science questions across the field of physics, chemistry, biology, mathematics, zoology, botany, designed to contain incorrect assumptions. Each question was fed to the language model GPT-4, and responses were recorded. The analysis involves categorizing the questions based on the type of logical fault and comparing the accuracy of the model's response.

Categorization: Questions are grouped together based on the type of error they contain.

Type of Fault: Questions involving incorrect physical assumptions such as resistance of a circuit being negative, had higher chances of getting a wrong response from these models.

Response Accuracy: GPT-4 very often failed to identify errors in questions which require a deep understanding of the physical laws, suggesting a gap in its training on fundamental physics concepts.

Pattern Identification: Look for recurring approaches in how the LLMs handle different categories of flawed questions.

Model Consistency: The model was consistent in its approach to all types of errors, typically attempting to generate a plausible answer based on incomplete or incorrect data, rather than identifying the fault.

EXPERIMENTS:

This project with the dataset that I have created uses transformer-based Large Language Models (LLMs) like xlnet and distilbert for the classification of faulty science questions across different disciplines. The key objective was to fine-tune these models in order to detect logical inconsistencies in questions and categorize them. This capability is critical in understanding how LLMs interprets and processes incorrect scientific information, which can lead to insights into improving LLMs logical reasoning abilities.

Data Preparation

This project makes use of a dataset specifically designed with science questions that contain wrong logical flaws. The questions are tagged with their discipline; hence this provides a multi-class classification challenge. Initial data loading involves the extraction of relevant columns and assurance of data integrity by removing any entries with missing information. The discipline labels are numerically encoded using Label Encoder from scikit-learn, converting categorical labels into a format suitable for model processing. The dataset is split into a training subset containing 80% and a test subset containing 20%. This is very important in model evaluation on unseen data for its generalization.

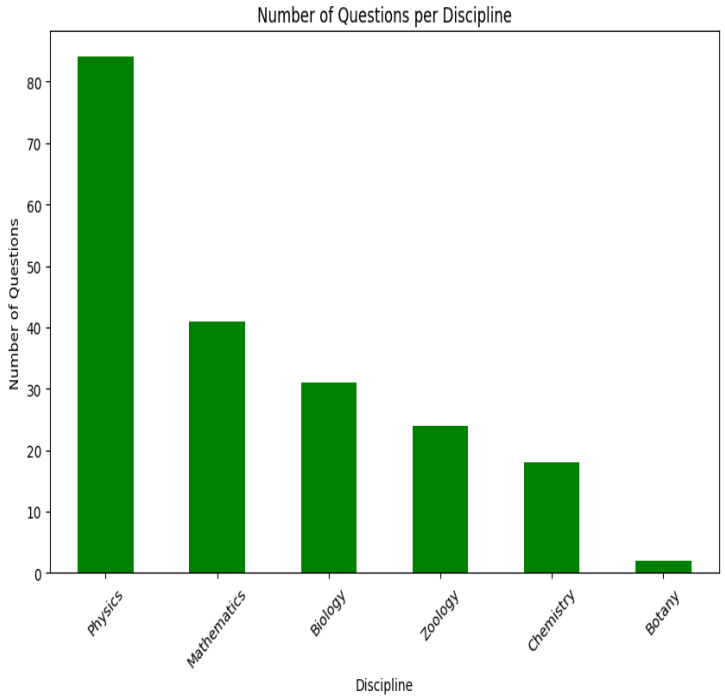
Model Setup and Training

xlnet and distilbert models are selected to perform this task because they are robust and have proven effectiveness in various NLP tasks. These models are pre-trained on large and offer a strong starting point for fine-tuning. Text data is tokenized using the respective model's

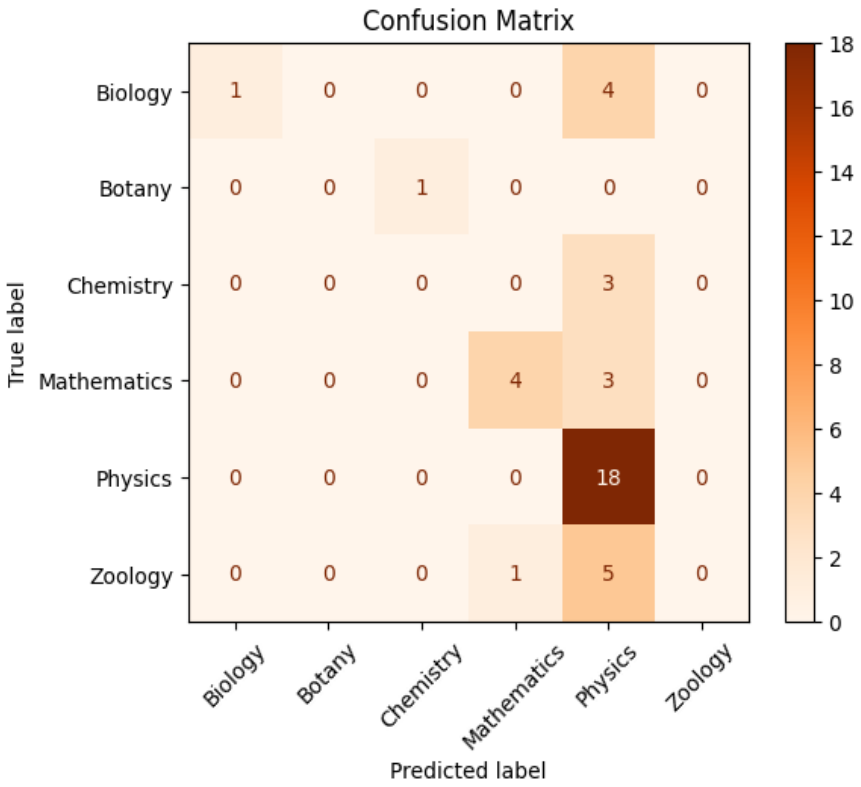
tokenizer. This is a process of converting text into a sequence of tokens that are input into the model. The tokenization includes padding and truncation to ensure uniform sequence length. Fine-tuning involves modifying pre-trained models for better performance with regard to the classification task. Training arguments are set to find the best learning rate, batch size, and number of epochs, considering training speed with model accuracy.

Training and Evaluation

Using the Hugging Face Trainer API, the models are trained on the tokenized training data enabling real-time logging of the process to monitor its progress. The models learn to associate the patterns in question text with the corresponding labels of logical inconsistencies. The models after getting trained are evaluated on the test dataset. Key metrics such as accuracy, precision, recall, and F1-score will be computed to assess each model's ability to correctly classify the faulty questions. The results show the performance of each model in identifying logical faults at different scientific disciplines by pointing out their strengths and weaknesses.



The bar graph above shows the distribution of faulty science questions across various disciplines. Physics is the most represented discipline in the dataset, with over 80 questions. Mathematics and Biology follow, with about 40 and 30 questions, respectively. Zoology, Chemistry, and Botany feature fewer questions, which could be a potential area for expanding the dataset to ensure a more balanced representation across all scientific disciplines.



The above confusion matrix depicts the performance of the model on classifying six scientific disciplines. It performs extremely well in identifying Physics with 18 correct predictions, while the same weaknesses with similar fields, Biology is often mistakenly classified as Zoology, and Chemistry and Mathematics are both very often misclassified as Physics. These errors suggest that the model has difficulty distinguishing between disciplines with overlapping features, therefore needing more training and distinguishing between features. Overall, the matrix illustrates the strengths and areas for refinement in the model.

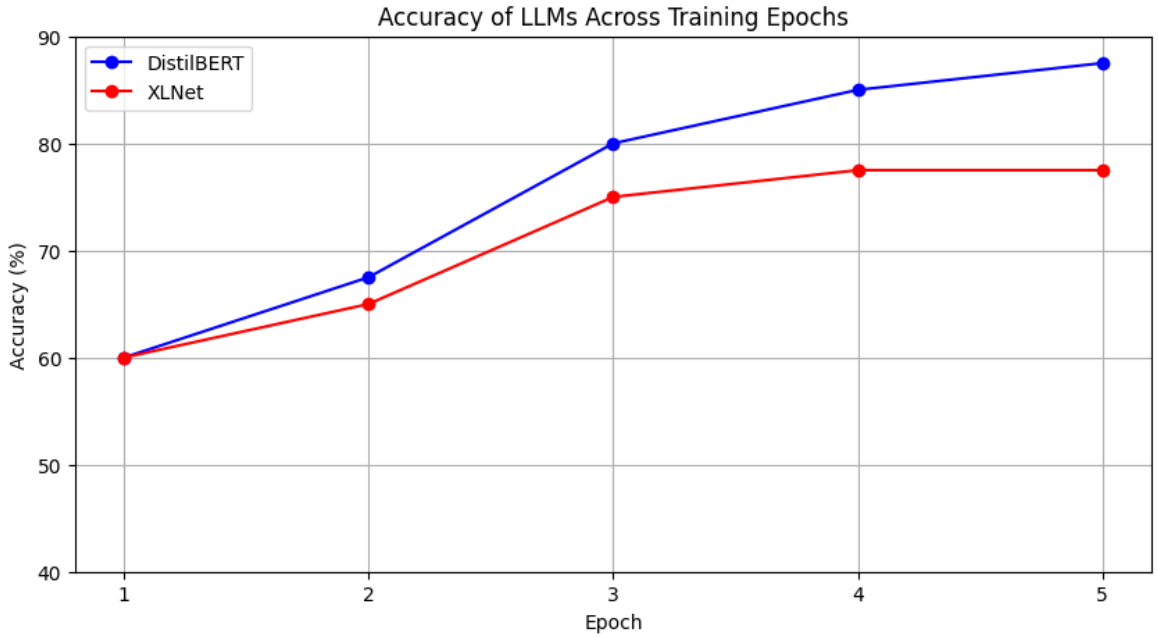
RESULTS:

LLM Model	Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1 score
DistilBERT	1	No log	1.3356	60.0%	41.32%	60.0%	47.31%
	2	No log	1.0483	67.5%	64.60%	67.5%	61.25%
	3	No log	0.8822	80.0%	75.57%	80.0%	76.35%
	4	No log	0.7895	85.0%	78.88%	85.0%	80.94%
	5	No log	0.7528	87.5%	88.31%	87.5%	85.79%
XLNet	1	No log	1.2283	60.0%	41.32%	60.0%	47.31%
	2	No log	0.8309	65.0%	60.43%	65.0%	58.12%
	3	No log	0.7261	75.0%	73.50%	75.0%	72.06%
	4	No log	0.6591	77.5%	71.36%	77.5%	73.92%
	5	No log	0.6361	77.5%	71.36%	77.5%	73.92%

The above table shows that the results for the experiments we have done on Distil BERT and XLNet:

Distil BERT showed significant improvements over all the metrics as it was trained. Whereas it started off in epoch 1 with an accuracy of 60.0% and an F1 score of 47.31%, showing the early stages of struggle to adapt to the task, in epoch 5 it jumped up significantly, reaching 87.5% in accuracy and 85.79% in the F1 score. The same improvements happened with precision and recall, meaning the model became more reliable and consistent in its predictions.

XLNet also improved, but never at the level of Distil BERT. From a starting point of 60.0% accuracy and an F1 score equivalent to Distil Bert's starting, XLNet increased to 77.5% by epoch 5, with its best F1 score at 73.92%. While these numbers do indicate growth, the precision and recall scores of the model remained lower than those of Distil BERT, indicating areas where XLNet could be further optimized to handle the classification task better.



OBSERVATION:

These results show that language models are good at processing well-structured questions but cannot easily reject flawed premises. This is because such a model would need to be trained on an error dataset so that it learns critical reasoning. Initial analysis of the dataset reveals several interesting patterns, many of the questions involve impossible scenarios, such as negative resistances. There's a recurring theme of questions that have missing critical information, especially in calculation-based queries. Conceptual misunderstandings, especially in areas such as special relativity and quantum mechanics, are common. The LLMs often give numerical answers to questions for which sufficient information to calculate an accurate answer is not provided.

CONCLUSION:

The behavior of large language models (LLMs) was explored as they interacted with my dataset with scientifically impossible questions. The investigation revealed that LLMs often attempt to provide solutions even when faced with scenarios that don't exist in real-world possibilities, highlighting a critical shortcoming. Notably, the results demonstrated that while LLMs occasionally recognized anomalies in questions involving clear physical impossibilities.

The results shown in this project underscore a critical challenge in applying the language models to educational and scientific contexts, their tendency to process and respond to queries without evaluating the logical consistency of the questions. Although models like ChatGPT 4.0 and Claude are capable of recognizing obvious physical impossibilities, they often fail to question the logical errors of mathematical problems. This suggests that solution generation might be overemphasized at the expense of problem validation. Significantly, the experiments showed that with explicit prompted, both models showed enhanced ability to discern and reject faulty premises, suggesting that these systems possess inherent capabilities for critical evaluation but may require explicit to engage these mechanisms effectively.

The fine-tuning of xlnet and distilbert for this classification task demonstrates the potential of transformer models to enhance LLMs understanding of logical consistency in scientific discourse. The study not only contributes to the field of LLMs in science education but also offers a pathway for further research into developing AI systems capable of deeper logical reasoning. This study underscores the need for language models to not only understand and generate human-like text but also to evaluate the validity of the content they are processing.

FUTURE WORK:

This research offers pivotal insights into the capabilities and limitations of large language models (LLMs) when addressing faulty scientific questions. The implications are broad, impacting the development of more robust error-checking mechanisms and enhancing LLMs ability to discern and clarify ambiguous queries. This is particularly crucial in contexts such as science education and automated tutoring systems, where the accuracy and reliability of content are paramount. To advance the utility and reliability of LLMs in academic and professional settings, future research should prioritize several key areas:

1. Expansion of Training Datasets: There is a critical need to broaden the diversity and volume of training examples that present faulty or illogical queries. By doing so, LLMs can better generalize and recognize these instances in real-world applications, enhancing their performance and utility. This effort would involve not only increasing the quantity of training data but also the complexity and subtlety of the logical errors included, thus preparing the models to handle a wider range of erroneous inputs.

2. Application of Findings to LLM Teaching Strategies: The insights obtained should be used to develop AI educational tools that go beyond mere knowledge delivery to foster critical thinking and problem-solving skills among learners. By utilizing LLMs that effectively identify and correct faulty questions, educational technologies can be designed to

engage students more deeply and encourage a more profound understanding of scientific concepts.

3. Automated Prompt Engineering: Further research into automated techniques for effective prompt engineering could significantly enhance the usability of LLMs without extensive human oversight. Such developments would enable AI systems to adaptively generate or refine their prompts based on the task user input, thereby improving interaction quality and model reliability.

By addressing these areas, future research can mainly focus on identifying the flaws in the question before trying to answer those questions.

REFERENCES:

1. *Taylor, L., & Jackson, M. (2023). A Comprehensive Overview of Techniques for Generating Text with GPT Models. arXiv preprint arXiv:2305.07890.*
2. *https://huggingface.co/*
3. *ChatGPT, GPT-4, Gemini 1.5, Claude*