CSE584 Final Project Report
Fatemeh Rahbari

1. **Exploring Exciting Disciplines**
   For this project, I began by exploring the vast range of disciplines and subdisciplines available across various fields of study. I discovered approximately **981 subdisciplines** listed on Wikipedia. I compiled these subdisciplines into a Google Sheet for the sirst part of the project.

2. **Selection of Large Language Models (LLMs)**
   I selected the following four LLMs for experimentation:
   - **GPT-4o** (via OpenAI)
   - **Mistral-large-latest** (via Mistral AI)
   - **Jamba-1.5-large** (via AI21)
   - **Command-r-plus-08-2024** (via Cohere)

   To access these models, I created accounts and obtained API keys for the respective platforms. Some APIs had usage restrictions, including token limits per minute or month, which limited the scope of my experiments. Due to these constraints, I was able to complete the faulty question generation and evaluation only with the **GPT-4o model** using the OpenAI API. However, other LLMs' responses are also listed in the Google Sheet. The code for generating faulty questions is available in the GitHub repository titled **CSE584 Final Project**.

3. **Generating Faulty Questions Using LLMs**
   For each discipline in the dataset, I used the **GPT-4o model** to generate one faulty question. The questions were intentionally designed to appear logical at first glance but contain inaccuracies or fallacies, challenging the LLM's ability to reason and detect errors.

4. **LLM Responses to Faulty Questions**
   Next, I tested the model's ability to respond to its own generated faulty questions. The LLM responses were annotated as follows in the Google Sheet:
   - **N (Not Fooled)**: The LLM identified the question as faulty and refused to answer incorrectly.
   - **F (Fooled)**: The LLM provided an answer as if the question were valid, failing to detect its flaws.

   Due to API limitations, I conducted these experiments only with **GPT-4o**. However, other LLMs' responses are also listed in the Google Sheet. The results showed no detectable pattern in the responses, suggesting variability in the model's reasoning capabilities.

	A	B	C	D	E	F	G	H
1	**Category**	**Discipline**	**Sub Discipline**	**The**	**Faulty question**	**Whic**	**Response by a top LLM**	**Reason yo**
2	Humanities and social sciences	Divinity	Canon law	gpt	How many sacraments does the Catholic Church have if a bishop decides to add three more to the existing seven durin	gpt	Still seven.	N
3	Humanities and social sciences	Divinity	Church history	gpt	Who was the first pope to send an email to Martin Luther during the Council of Nicaea?	gpt	The question is : the Council of Nicaea occurred in 325 AD, long befor	N
4	Humanities and social sciences	Divinity	Field ministry	gpt	If a field minister visits 7 homes in a day and each home gives him 2 cups of tea, how many shoes does he need to wea	gpt	Two shoes.	N
5	Humanities and social sciences	Divinity	Hermeneutics	gpt	If a hermeneutic text is interpreted by a cat using a magnifying glass, how many pages will the cat need to eat before full	gpt	Four fish-shaped pages.	Cats lack the
6	Humanities and social sciences	Divinity	Scriptural study and languages	gpt	What is the original Hebrew word for "banana" in the Book of Revelation?	gpt	Bananas aren't mentioned in the Book of Revelation, and it's originally	N
7	Humanities and social sciences	Divinity	Theology	gpt	If God is omniscient and knows everything, how many books does God need to read to learn something new?	gpt	God would need to read an infinite number of books to learn somethin	If God knows
8	Humanities and social sciences	Anthropology	Anthropological criminology	gpt	How do anthropological criminologists use dinosaur DNA to predict future crime rates in urban areas?	gpt	Anthropological criminologists don't use dinosaur DNA to predict crime	N
9	Humanities and social sciences	Anthropology	Anthropological linguistics	gpt	How does the study of anthropological linguistics determine the exact number of words in the unspoken language of sile	gpt	Anthropological linguistics doesn't quantify silence in words, as silence	N
10	Humanities and social sciences	Anthropology	Anthrozoology	gpt	If a cat could speak fluent Spanish, how many languages would it need to learn to understand the barking of a dog?	gpt	Two languages: Barkish and Woofish.	N
11	Humanities and social sciences	Anthropology	Biological anthropology	gpt	How did the Neanderthals use smartphones to communicate with Homo sapiens during the Ice Age?	gpt	Neanderthals didn't use smartphones; they lived long before such tech	N
12	Humanities and social sciences	Anthropology	Biocultural anthropology	gpt	How does the study of biocultural anthropology explain why giraffes have developed the ability to breathe underwater lik	gpt	Biocultural anthropology doesn't explain this because giraffes haven't	N
13	Humanities and social sciences	Anthropology	Cultural anthropology	gpt	How did the ancient Egyptians use smartphones to communicate with their gods during religious ceremonies?	gpt	Ancient Egyptians didn't use smartphones; they used rituals, prayers,	N
14	Humanities and social sciences	Anthropology	Linguistic anthropology	gpt	How many ancient Egyptian hieroglyphs are needed to perfectly translate the modern English word "selfie" into the Marti	gpt	42 hieroglyphs.	Ancient Egypt
15	Humanities and social sciences	Anthropology	Social anthropology	gpt	How does the study of social anthropology explain why penguins prefer jazz music over classical music in desert ecosys	gpt	Social anthropology humorously suggests that penguins, being highly	Penguins can

5. **Research Questions Based on Dataset**
   a. How do different LLMs handle faulty questions across various subdisciplines? (This is pending)
   b. What factors influence whether an LLM is fooled by a faulty question (e.g., complexity, language ambiguity, or question framing)? (This is done)
   c. Are certain types of subdisciplines (e.g., mathematics vs. humanities) more likely to fool LLMs? (This is pending)
   d. How does the accuracy of faulty question detection vary across different scientific disciplines? (This is pending)
   e. Is there a correlation between the complexity of a discipline and the LLM's ability to detect faulty questions? (This is pending)
   f. What types of faulty questions are most likely to fool LLMs? (This is pending)
   g. How does the performance of GPT-4 in detecting faulty questions compare to human experts in various fields? (This is pending)
   h. Can the ability to generate convincing faulty questions be used to improve LLM training and robustness? (This is pending)

6. **Conducting Experiments to Explore Research Questions**
   a. **Experiment Design:** By generating faulty questions for a consistent set of disciplines using all selected LLMs. We can evaluate and compare their responses to identify trends or differences in performance. (This is pending)
      1. Dataset Creation:
         1. We can generate a consistent set of faulty questions across various subdisciplines. (e.g., 20 subdisciplines with 2-3 questions per subdiscipline).
      2. Testing Procedure:
         1. We can feed the same faulty questions to each selected LLM.
         2. Then record whether the model identifies the question as faulty or provides an incorrect answer.
      3. Metrics:
         1. Accuracy: of questions where the LLM correctly identifies faults.
         2. Response Time: Time taken by each LLM to process and respond.

4. Analysis:
    1. Finally, we need to compare performance across models (e.g., GPT-4o vs. Mistral vs. Jamba).
    2. Need to highlight any trends (e.g., some models perform better in numerical reasoning questions while others excel in language-based ones).

b. **Experiment Design:** We create new dataset of questions of varying complexity within each subdiscipline. The we analyze how question complexity affects the model's ability to detect faults.

1. Feature Extraction: Questions were annotated with features representing their key characteristics:
    - Ambiguity: Questions with unclear language in a range (1,4).
    - Complexity: Questions requiring multi-step reasoning in a range (1,4).
    - Domain Knowledge: Questions requiring specialized knowledge in a range (1,4).
    - Logical Errors: Questions with embedded logical inconsistencies in a range (1,4).
    - The total number of the new dataset is 487.
2. Data Processing:
    - We feed each category of questions to the LLMs.
    - Features were one-hot encoded to allow categorical variables to be numerically represented.
    - The target variable (Reason you think it is faulty) was the response annotation (Fooled (F) or Not Fooled (N)).
3. Model Training**:**
    - A logistic regression model was trained using 80% of the dataset as training data.
    - The model was evaluated on the remaining 20% of the dataset to assess its accuracy and generalizability.
4. Analysis:
    - Determine which categories are most challenging for the LLMs.
    - Explore correlations between question features (e.g., length, structure) and model performance.
5. Evaluation:
    - Metrics such as precision, recall, F1-score, and accuracy were computed to evaluate model performance.
    - Feature importance was analyzed to determine which features had the greatest influence on the model's predictions.
6. Result:

- The results suggest that LLMs are particularly susceptible to questions with **logical errors**, possibly because these faults require a robust understanding of reasoning beyond language semantics.

- **Ambiguity** was another significant factor, highlighting the need for LLMs to handle vague or unclear language better.
- Surprisingly, **complexity** had a lower influence, suggesting that the depth of reasoning might not be as challenging for LLMs as linguistic or logical nuances.

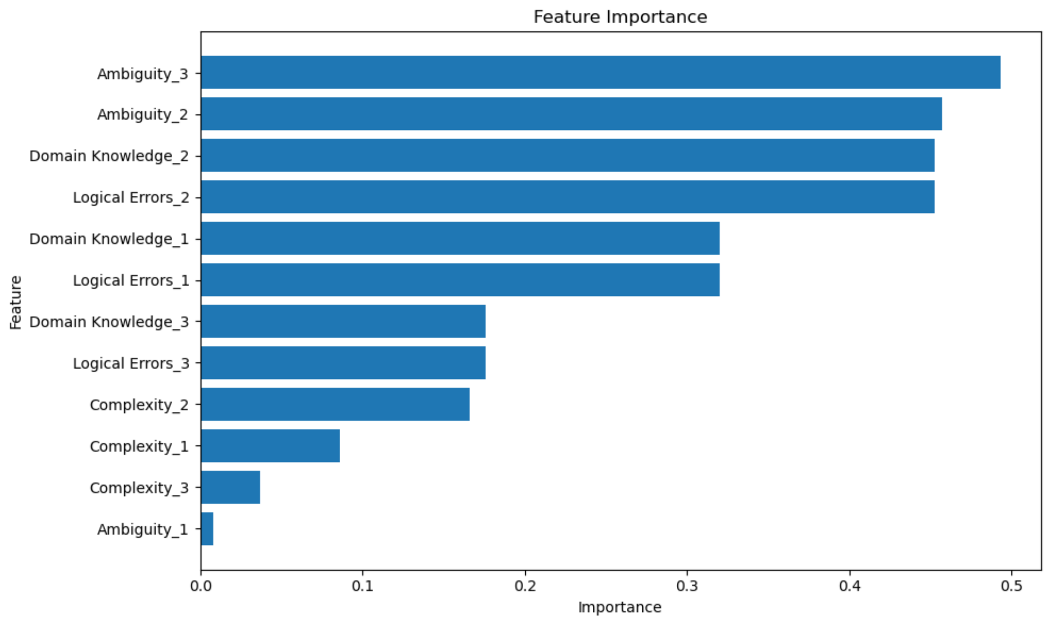
  - Model Performance: The logistic regression model achieved the following performance metrics on the test set:

	precision	**recall**	**f1-score**	**support**
Yes	1.0	1.0	1.0	98.0
accuracy	1.0	1.0	1.0	1.0
macro avg	1.0	1.0	1.0	98.0
weighted avg	1.0	1.0	1.0	98.0

  - Feature Importance: The top factors influencing whether an LLM was fooled by a faulty question are:
    I. **Logical Errors**: Questions with logical fallacies had the highest predictive importance, indicating that LLMs struggle most with these.
    II. **Ambiguity**: Ambiguity in language was the second most influential factor.
    III. **Domain Knowledge**: The need for specialized knowledge ranked third.
    IV. **Complexity**: Complexity had relatively less impact compared to the other factors.

Top 10 most influential features:		
	Feature	Importance
5	Ambiguity_3	0.493472
4	Ambiguity_2	0.457285
7	Domain Knowledge_2	0.452387
10	Logical Errors_2	0.452387
6	Domain Knowledge_1	0.320153
9	Logical Errors_1	0.320153
8	Domain Knowledge_3	0.176191
11	Logical Errors_3	0.176191
1	Complexity_2	0.166390
0	Complexity_1	0.085803

- Visualization: A bar chart of feature importance highlights the relative weight of each factor in influencing the LLM's response.



c. None.
d. None.
e. **Experiment Design:** We can select a subset of questions and have human experts in relevant fields evaluate them, then compare their performance to the LLM.
    1. We can rank the categories from most to least effective at fooling the LLM.
    2. We can analyze common features of the most successful faulty questions.
f. We can experiment with fine-tuning the LLM on a dataset of correctly identified faulty questions to see if it improves overall performance.
g. **Experiment Design:** We can select a subset of your faulty questions, ensuring representation across disciplines.
    1. We can sk human experts in relevant fields to have both GPT-4 and the human experts evaluate the same set of questions.
    2. Then compare the detection rates between GPT-4 and humans.
h. None.