Richard Ma CSE 584 Final Project

1. Introduction and Methods

A set of 66 faulty science questions was presented to an LLM to observe patterns in the LLM's ability to detect and correctly respond to the faulty nature of the question. For all questions, the GPT-4o mini LLM (henceforth 'the LLM') was used for consistency. A full list of questions and responses can be found in the appendix.

The questions vary by multiple parameters, listed below. This study will examine how the interaction of these parameters within faulty questions influences the LLM's response:

- The language in which the question was asked: Every question was asked in both English and Spanish. Questions were originally written in English and manually translated to Spanish. The original meaning and wording were preserved to the fullest extent possible..
- Discipline: Questions were posed within four disciplines: solid mechanics, materials science, linguistics and history
    - History and linguistics questions can be further distinguished by their relevant year. Due to the relative sparseness of the dataset, datable events are grouped into broad time periods–antiquity, medieval, early modern, 19th century and modern day. Modern (non-extinct) languages will be assigned the current year (2024).
- Question faultiness category: Questions differed in the manner by which they were faulty. For example, the largest faultiness category consists of questions concerning nonexistent items (The 'Does not exist' category). Other types of faulty questions are grouped simply into the 'other' category. An example of a "Does not exist" question is the following: What was the main argument of William F. Chesterfield's speech before parliament in 1839? The faultiness of the question arises from the fact that William F. Chesterfield and his speech did not exist in reality.

The LLM's response to these questions is divided into correct, incorrect and partially correct subcategories. Correct responses correctly acknowledged and responded to the fault present in the question. Incorrect responses did not acknowledge the fault and proceeded with the response as if the fault aligned with reality. Partially correct responses either recognized the fault or did not fully respond to the question as if the fault aligned with reality, although they responded incorrectly or inadequately in some other capacity.

- As an example of a correct response, when asked for the etymology of the nonexistent word 'ornasalum' in Classical Latin, the LLM responded that this word is not attested.
- As an example of a partially correct response, when asked how to inflect the Russian noun 'дверь' into the vocative case, the LLM correctly responded that this word could not be further inflected into the vocative. However, it incorrectly gave as a justification that this noun ends in 'ь'. In reality, the noun's ending and class are completely irrelevant to vocative inflection; the correct answer is that the vocative case is extinct in modern Russian with the exception of extremely limited religious contexts and colloquial usage, both of which do not apply to 'дверь'.
- As an example of an incorrect response, the LLM was asked (in Spanish) for the four parameters of the Swift hardening law (solid mechanics). The LLM incorrectly responded with four hallucinated parameters, when in reality the Swift hardening law only contains three parameters
- For brevity, incorrect and partially correct answers will be combined into a single category in plots when appropriate, and some categories divisions will have no partially correct responses. The full dataset with all information can be found in the appendix.

2. Broad categories:

An analysis will first be conducted of the response of the LLM when considering broad categories, or those to which every question in the dataset can be assigned. These broad categories are question language (English or Spanish), discipline and question faultiness category. The breakdown of these categories can be seen in **Figure XX**. As seen, the distribution of questions by discipline is very roughly even. Within each discipline except for materials science, the majority of questions are of the 'does not exist' category.

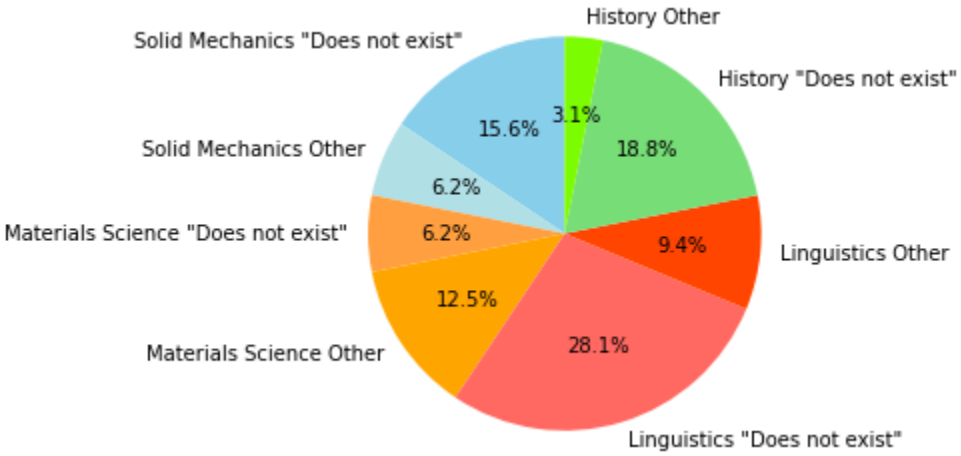

Figure XX: Breakdown of questions by category. Not shown here, the query language in every sub-category is split perfectly evenly between English and Spanish

As shown by **Figure XX**, the effect of query language overall had little effect on LLM response. Considering English-Spanish question pairs, the LLM responded the same way to 82% of the two paired questions and differently to 18%. Within these pairs with different responses, the English version of the question received a better response (correct better than partially correct, in turn better than incorrect) 67% of the time while the Spanish version received a better response 23% of the time. This possibly indicates an improved ability to detect faulty queries posed in English over those posed in Spanish, although the quantities of data are not sufficiently statistically significant to demonstrate this.

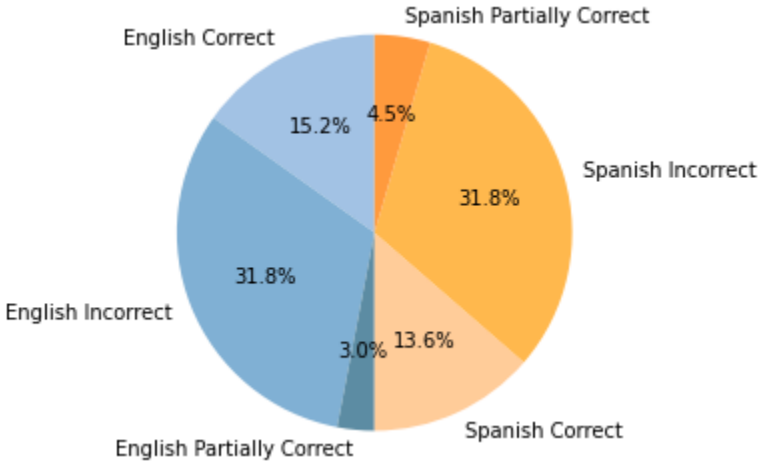

Figure XX: The effect of query language on LLM response

As seen in **Figure XX,** a clear difference in response ability can be seen based on discipline. A majority of linguistics questions are answered correctly, a minority of history and materials science questions are answered correctly, and no solid mechanics questions are answered correctly. Considering Figure XX and Figure XX, no clear correlation can immediately be seen between response ability based on discipline and the question categories ("Does not exist" and "Other"). This effect likely stems from the richness of information with which the LLM was trained. Especially regarding the existence and etymology of words, the entire body of training data in a particular language, regardless of discipline/subject matter and not limited to specific academic linguistic literature, could have implicitly imparted linguistics knowledge, thus explaining the relative excellent ability of the LLM to detect faulty linguistics questions . History is also a well-documented subject compared to materials science and solid mechanics.

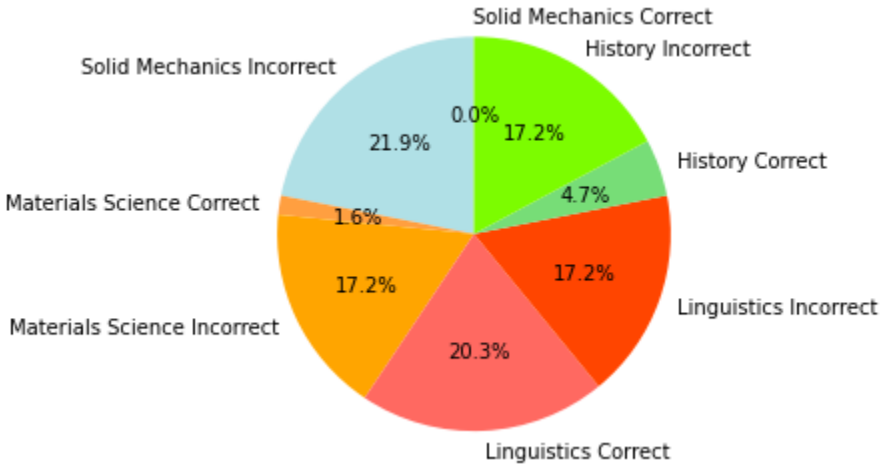

Figure XX: The effect of query language on LLM response. "Incorrect" and "Partially correct" categories are combined. No faulty solid mechanics questions were answered correctly.

As seen in Figure XX, there was no significant difference in LLM response ability to "Does not exist" questions and to other types of questions. The former entailed a correct response rate of 30% while the latter entailed a correct response rate of 28%. As such, the question type distinction likely has little interaction with the other observed category effects.

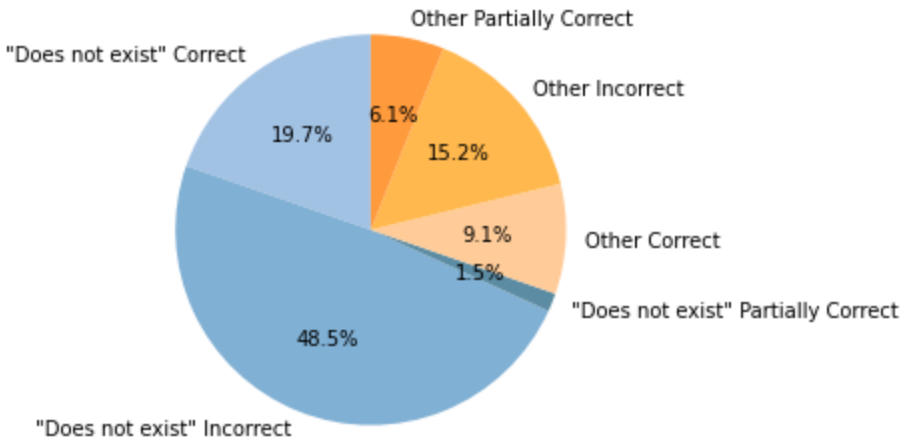

Figure XX: *The effect of query type on LLM response*

3. Inter-category analysis

The effects of the three broad categories having been analyzed, inter-category effects will now be explored. History and linguistics responses may be broken down by time period, as seen in Figure **XX**. At face value, the antiquity and modern time periods have the greatest percentage of answers correctly answered. Considering the nature of questions asked within each time period, a strong correlation can again be hypothesized between richness of training data and accuracy of faulty response detection. All but one of the antiquity questions concerned late republican Rome and the classical Latin language, both of which are richly documented. Every faulty question concerning Rome, in fact, was correctly answered.

Similarly, the modern linguistics faulty questions can be divided into two categories—those concerning the modern Russian language and "Does not exist" questions concerning invented languages. No faulty questions concerning the Russian language were answered incorrectly—one was answered partially correctly, and all others were answered correctly.

Medieval and early modern questions mainly concerned regions and languages which, compared to the modern world and late republican Rome, are much more poorly documented.

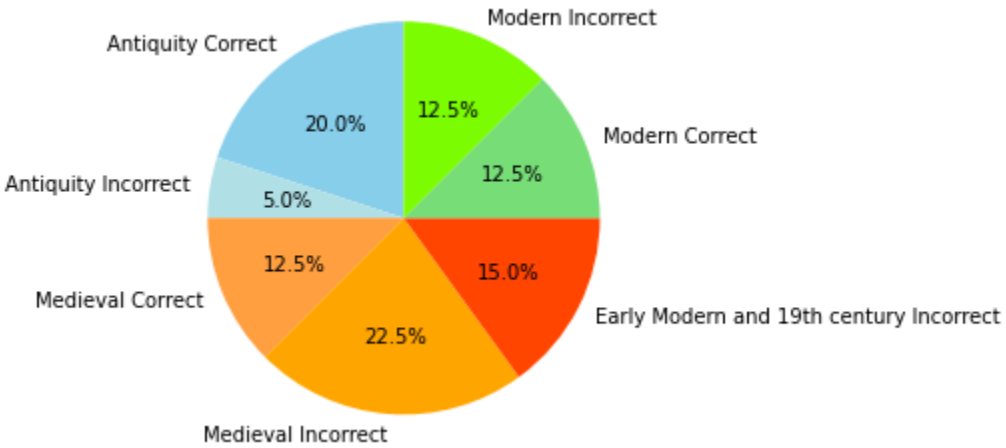

Figure XX: The effect of time period on LLM response in history and linguistics questions. No early modern or 19th century questions were answered correctly. Every question in the 'modern' time period is linguistics related. "Incorrect" and "Partially correct" categories are combined.

4. Conclusion

In this study, a set of faulty questions was prepared for testing of the GPT-4o mini LLM. Each question is describable by the language in which it was written (English or Spanish), academic discipline (solid mechanics, materials science, history, linguistics), and faultiness category ("Does not exist" questions, others). History and linguistics questions were further describable by time period. Considering the availability of data, high-dimensional interactions between these categories was not feasible.

Across the overall dataset, little effect on the LLM's ability to respond to faulty questions was observed due to question language and faultiness category. A strong effect was observed, however, from academic discipline. The majority of linguistics faulty questions were answered correctly. A minority of history and materials science faulty questions were answered correctly, and no solid mechanics questions were answered correctly. It is hypothesized that this is due to a strong correlation with the richness of the LLM's available training set–that almost all of the training data, regardless of subject matter, could have implicitly served as linguistics training data.

Within the history and linguistics dataset, the LLM was able to disproportionately respond correctly to faulty questions related to late republican Rome/Classical Latin and modern Russian. In these cases, no questions were answered incorrectly. It is hypothesized that this is also due to a strong correlation with available training data.

In short, response ability to faulty questions has been shown to be strongly affected by the subject matter of the question, but not by the manner in which the question was asked or by the nature of the question's faultiness.

5. Appendix

Dataset representations:
Version submitted 11/30/24:
CSE584/Richard Ma Final Project CSE584 Dataset .xlsx at main · marichard123/CSE584

A richer version used to create plots:
CSE584/final_project_csv_dataset_2.xlsx at main · marichard123/CSE584

Text document with full LLM responses to all faulty questions:
CSE584/Richard Ma GPT40mini Full Responses.txt at main · marichard123/CSE584