Patterns of success and failure: Analysing Large Language Models in Question Answering in Exam Contexts

FELICIA BURLACU, University of Twente, The Netherlands

Abstract

Large Language Models (LLMs) have come to be the spotlight of the general public in the recent years, capturing the attention of people among different industries besides tech. LLMs are artificial intelligence (AI) models designed to interact with, process, generate and analyze human language on a large scale using deep learning techniques. It was through their high performance for complex human-level tasks like sentiment analysis [1], text summarization [2] and question answering [3], that these models gathered so much attention. However, despite their impressive results and substantial computational abilities, LLMs are not without drawbacks like biases and lack of interpretability in decision-making. But even with these challenges in mind, there are media articles published with titles saying 'GPT-4 beats 90% of the lawyers trying to pass the bar' [4], or 'Chat-GPT passes the Radiology Board Exam' [5] citing papers with recent research on GPT performance on such examinations [6] [7]. One reason for such evaluation methods would be to apply the LLMs in the context of real world problems in order to show their applicability but also to make these assessments of performance accessible for the general public without much technology background. This research aims to employ a few-shot learning approach in order to identify and analyze patterns in the performance of LLMs, specifically in the context of answering exam-type questions.

Additional Key Words and Phrases: Large Language Models, Natural Language Processing, Question Answering, GPT-3.5

1 INTRODUCTION

In recent times, LLMs have become a leading technology in the field of AI, specifically in the area on Natural Language Processing (NLP). These models are designed to process, interpret and generate human-like text through deep learning techniques, including using architectures like Transformers. Proposed by [8], Transformers introduce the new concept of self-attention, which allows to weigh the importance of every word in a sentence regardless of its positional distance. This makes it possible for LLMs to excel at a variety of tasks such as sentiment analysis, where the model detects the emotional tone behind a text [1]; named entity recognition which involves identifying and categorizing key elements from text into a given set of categories [9], and many more [10], [11]. Despite their versatility, LLMs face significant challenges. These include a tendency to replicate and amplify the biases present in the training data, high environmental costs due to the energy demands of training and updating such expansive models [12], hallucinations caused from the data, training or inference [13], lack of transparency in how the outputs are generated which makes the interpretation of their decisions difficult [14], [15].

Focusing on the task of question answering, especially in the context of exams, introduces additional layers of complexity. One factor that complicates the matter is that the format of questions might vary a lot within just one exam - ranging from multiple choice that test the student's ability to recognize correct answers, to open-ended questions that require extensive explanations, summarization of concepts, or creative and critical thinking. This presents a challenge for the LLM as it must adapt not only to the content of the question, but also generate precise and relevant answers in a very specific style.

The motivation for this research that focuses on exams, because by their nature exams depict real-world problems, as the questions they are composed of are designed not only to test knowledge, but also the capability to apply it in varied and complex scenarios. This is representative of how tasks in the professional fields look like, requiring to demonstrate a varied set of cognitive abilities. Moreover, focusing on exams allows to study the LLM's capabilities in a controlled, yet complex environment by testing the model's ability to navigate through different categories of questions that represent different real-world problems. This serves to close the gap between theoretical AI and practical and applicable skills, offering insight into how the LLMs can be further improved and utilized in various scenarios where decision-making and problem solving are of high importance. The main goal of this research is to explore and uncover patterns in LLM performance in exam contexts, identifying specific types of questions in which the models do best or worst. By employing a few-shot approach, we adapt the LLM to the various characteristics of exam questions with minimal extra training data in order to assess its performance in a setting with reduced context given.

1.1 Research Question

While there has been extensive research done in the field of AI and LLMs and a large number of papers explored their ability to pass examinations [6], [7], [16]–[18], there is a notable gap when it comes to understanding their application and performance in the context of question answering during exams especially when it comes to the patterns (of questions) and factors that influence the performance metrics. This paper aims to address this gap by conducting a thorough review of LLMs' performance across various exam contexts, exploring both their strength and limitations. It will examine how these models handle the unique challenges posed by the diverse formats and complexity levels of exam questions. The problem statement leads to the following research question:

Which aspects of the question structure influence the performance of LLMs in answering exam-style questions?

TScIT 41, July 5, 2024, Enschede, The Netherlands

Question Type	Multiple-Choice	Short Answer	Open-ended
Short (<50 words)	7	8	-
Medium (50-100 words)	8	7	-
Long (100< words)	8	7	-
Factual	7	7	-
Analytical	7	7	-
Synthesis-based	-	-	17

Table 1. Distribution of questions from the dataset

This can be answered with the following sub-questions:

- *RQ1: How do variations in question format (e.g., multiple choice, short answer, open-ended) affect the LLM response?*
- *RQ2: What is the effect of question length on LLM response quality?*
- *RQ3: How do variations in question types (e.g., factual, analytical, synthesis-based) influence LLMs response?*

2 LITERATURE REVIEW

When exploring the performance of LLMs in a few shot setting, there are several research areas that are relevant to our research and would stand as background for our further study.
After the introduction of the Transformer architecture [8] followed the introduction of such LLMs as BERT by Google [19], LLaMa by Meta AI [20], Gemini by Google[21], ChatGPT by OpenAI [22].

While these have pushed the boundaries in NLP, these advancements also highlight evaluation challenges. [23] discussed how due to these models reaching or surpassing human-level performance, traditional benchmarks may no longer provide meaningful insights, due to the 'superhuman abyss' where the human yardstick becomes insufficient. Therefore, the evaluation of such transformative models needs a reconsideration of AI benchmarks to encompass multi-dimensional scales and metrics that can capture the breadth and depth of AI advancements.
There have been multiple studies in the field focusing on evaluating the performance of LLMs in exam contexts [6], [7], [16]–[18], [24]–[30], . For instance, [29] includes the study of the performance of LLM integrated chatbots: GPT-3.5 and 4, Google Bard and Bing Chat in solving ophtalmology fellowship exams. [26] showed that GPT-4 outperforms humans and GPT-3.5 in an ophtalmology self-assessment program, highlighting its performance in multiple-choice question formats in medical fields. Another study [24] includes insights into GPT-4 performance on biomedical science exams, more specifically concluding about its great performance on fill-in-the-blank, short answer and essay questions, and poor results on tasks regarding figures and requiring a hand-drawn answer. Hallucinations and flagged plagiarism were also identified for some of the answer-sets. [25] also emphasizes the the clinical reasoning of GPT-4.0 as well as the high accuracy of the answers. However, the authors encourage users to take special caution as the elaboration in erroneous responses is comprehensive and seems very accurate. OpenAI states on the page of ChatGPT-4 [31] that it scores in the 90th percentile for the Uniform Bar Exam. [32] critically examines this claim. Some of the arguments for why evaluation on exams like the Uniform Bar Exam include: misleading percentile rankings, overstated capabilities and lack of disclosure regarding essay questions, skewed results due to comparison of GPT results to a subgroup of test-takers who usually perform lower, the performance scores were not derived according to the standards typically used in bar exams. Another important aspect discussed in the paper is the fact that ChatGPT operates in an 'open book' context which is not realistic and presents as an unnatural advantage, as students take bar exams in closed-book conditions. [33] discusses and compares the capabilities of language models in different contexts, in fine-tuning and few-shot approaches. As some of the previously mentioned papers that assess the performance of LLMs in exams [25] use a fine-tuning approach it is relevant to point out that research [34] shows that due to the narrow, task-specific dataset that models are trained on, the risk of learning spurious correlations increases. Additionally generalization issues, over-fitting for a specific task were also discussed as drawbacks of fine-tuning. The paper shows that GPT-3 shows exceptional performance in a few-shot setting, outperforming in certain cases the fine-tuned state-of-the-art models.

3 METHODOLOGY

3.1 The LLM

The literature review showed that a large number of articles focus of the performance of OpenAI's GPT models and report on the good results it obtains in exams. Therefore it was chosen to follow this path and research about its few-shot capabilities in exam context and what impacts the performance. In order to employ the model in a few-shot setting in an efficient way, the LangChain library was used for accessing the GPT-3.5-turbo model and establishing prompt templates. The 3.5-turbo model is a fast, but inexpensive model from OpenAI's library which is intended for simpler tasks, therefore as we are using a few-shot technique, this version is suitable. Its architecture follows the previous GPT-3 model - an autoregressive model based on a Transformer architecture employing self-attention techniques. It has alternating dense and locally banded attention patterns in the layers of the transformer. Traditionally, Transformer-based models like GPT-2 and BERT make use of the "dense" or full attention mechanisms where each token in the sequence attends to every other token. In the case of GPT-3 and later models, the alternating dense and sparse locally-banded layers, meaning that in some of the layers the token will only attend to tokens that are nearby, thus creating the so-called "bands of attention". As a consequence, despite reducing computational costs by using this technique, its aim is to maintain and, in some cases, enhance the performance of the model by focusing the computational resources on the most relevant parts of the sequences.

3.2 The Dataset

One of the challenging parts of this study was gathering the necessary data to form a dataset. To address this, the dataset was manually constructed from the available examinations from the University of Twente (UT). One strong motivation for this was the necessity to consult domain-experts at the UT on open-ended, synthesis-based questions. Consequently, by sourcing the data from UT examinations, it was easier to match the questions requiring expert evaluation and the available researchers and professors at the university in order to get the required expertise. As a result, the final dataset contains 69 questions with answers divided in categories. Augmentation techniques like paraphrasing questions in order to increase the dataset robustness were not employed as while the number of question examples in the dataset would increase, the diversity would decrease. Additionally, it makes the management of the data much more complex in order to ensure no overlap of the example set of questions and the test question given. Without effective management, this could lead to skewness of the results as scenarios where the example set of questions are "giving away" the answer might happen. For instance, when the paraphrased version of a question from the example set is actually given as the test question for that iteration, the model would reproduce the answer from the example set rather than give a newly generated response. Based on the formulated research questions, the questions were divided into several categories in order to attempt to discover performance patterns. Based on RQ1, three categories were created: multiple-choice, short answer and open-ended, as to reflect the prevalent exam formats at the University of Twente. To answer RQ2, the questions were divided into short (less than 50 words), medium (50-100 words), long (more than 100 words). Lastly for RQ3, analytical, factual and synthesis-based questions were analyzed. Table 1 shows the distribution of questions from the database into the aforementioned categories.

3.3 Prompt Engineering

One of the effective way to interact with LLMs are prompts - instructions that dictate the guidelines and the specific requirements that the generated output has to corresponds to [35]. Few-shot techniques aim to guide the LLMs into generation of outputs according to desired format by the means of providing a small set of few-shot examples [33]. A five-shot technique was employed for the majority of the categories. From the question pool of the specific category, five questions with answers were randomly selected as examples and one random question for testing. This process was repeated seven times, each iteration acting as a fold in a cross-validation process. Because the LangChain library was used to access the GPT, the default function to retrieve its response to the given prompt does not include chat memory, as the implementation of Conversational Chain does. This enabled an effective iteration of retrieving answers for different folds without the need to restart the kernel and start with fresh memory. Additionally, the rotation of each of the question from an example-question and a test-question ensures that the performance metrics are not influenced by the fact that some questions might be inherently easier or harder than others.
It was decided to construct a prompt only containing the five questions with answers (the example set) and the unanswered question for the categories of MC and short answer question categories. There are several reasons for this: firstly, because these question formats do not require extensive answers, but extremely short answer texts, containing a single letter (in the case of MC questions) or a sentence, math formula or number (in the case of short answer questions). Secondly, for the goal of reducing inconsistent and biased answers that might be introduced by a persona-based prompts, and encouraging factual responses as required by these question formats. Open-ended questions were approached differently, as they require longer answers with a wider variety of responses that would be considered as correct. Additionally, the available exam solutions rarely include formulated answers to such questions, and rather include guidelines for grading. Thus, it was decided to apply a different approach for generating answers and their evaluation. A zero-shot approach was used, where the LLM was given the question and the corresponding instructions specifying the domain/course that the question belongs to, and the study programme. As this format requires long answers from students focusing their expertise in a specific field, the style and tone of the answer also is taken into consideration when evaluating responses. Additionally, employing a persona-based prompt with a minimal description of the role that needs to be employed, guides the LLM into giving an answer that includes more relevant to the field insights, making the answer more realistic, while also reducing the chances of overfitting to a persona encouraging answers that are both informed by the persona but also grounded in factual knowledge. The answers were then given to UT teaching assistants, professors and researchers for evaluation as part of a small user study.

3.4 Evaluation

In order to properly evaluate the generated responses for different question types, different evaluation methods were used.

Despite the fact that precision, recall and F1 score are common used evaluation metrics, evaluation on these metrics is only possible where multiple answers could be correct. Thus, they are not relevant in the use case of analysing the performance of LLMs on multiple-choice questions, as this question type only accepts one selection which is correct, and only positive (correct) and negative (incorrect) answers can be specified. Further classification of answers into a confusion matrix with True Positives (TP), False Positives (FP), True Negatives(TN) and False Negatives(FN) is possible only per class, or per answer choice (A, B, C, D, etc.), however these insights are not useful as attributing a class (letter corresponding to the choice) solely depends on the order in which the choices are presented. Due to this, it was decided to use accuracy as an evaluation metric, providing the proportion of questions for which the model selected the correct answers out of the total numbers of presented test questions.

Question Type/Length	Short (<50 words)	Medium (50-100 words)	Long (100<words)
Multiple-choice	57.14%	71.42%	42.85%
Short Answer	57.14%	42.85%	14.28%

Table 2. Accuracy of LLM responses by question type and length.

Question Type/Format	Factual	Analytical
Multiple-choice	71.42%	42.85%
Short Answer	57.14%	42.85%

Table 3. Accuracy of LLM responses by question type and format.

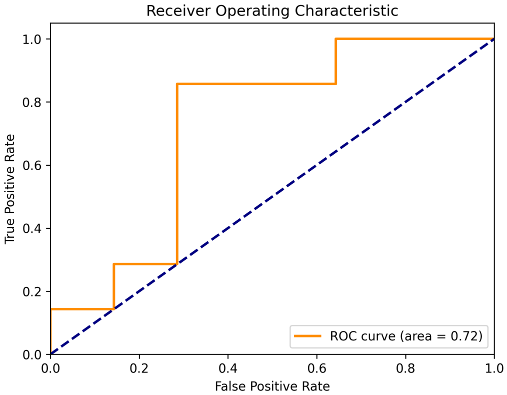

Fig. 1. ROC Curve

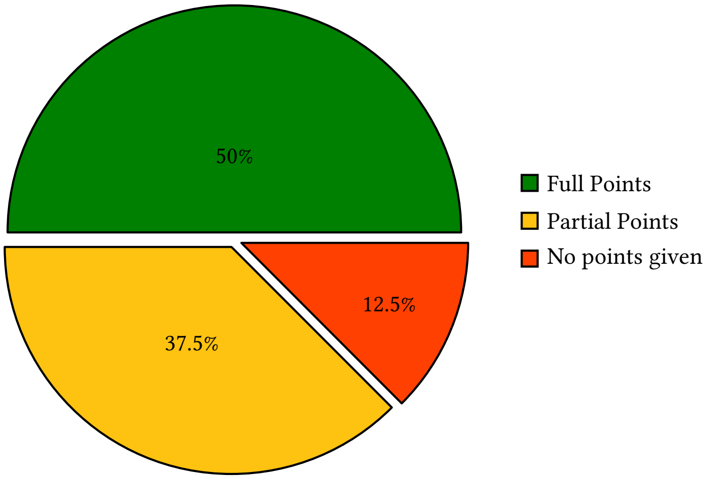

Fig. 2. Points scored

For short answer questions, BLEURT score [36] was used to measure to what extent the generated output reflects the meaning of the reference answer, as the correct answer can be formulated in a multitude of ways and still preserve the same meaning. It is an automated machine learning based metrics which uses the BERT model which is pre-trained on a public collection of ratings (WMT Metrics Shared Task dataset) and a list of user-given ratings to generate embeddings of the text pairs (test text, reference text) that are used to predict how close the text matches the reference, imitating human judgement [36]. By setting a threshold (t=0.537), the generated answers were categorized as Positive (above the threshold, therefore corresponding by meaning to the reference answer according to the BLEURT score) and Negative (below the threshold and do not convey the same meaning as the reference answers according to the BLEURT score).

The process of setting the threshold to an optimal value included gathering a set of generated examples along with their scores and their binary scores according to human judgement (1 represents acceptable answers - where the generated text is similar to the reference answer available; 0 represents unacceptable answers - the generated text is not similar to the reference answer available), generating a Receiver Operating Characteristic (ROC) curve shown in Fig. 1 to analyze the trade-offs between true positive rates and false positive rates for different thresholds and choosing a threshold that balances sensitivity(correctly identifying acceptable answers) and specificity (correctly rejecting unacceptable answers). Due to the variability of wording, structure and detail of answers that convey the correct response, it is not possible to categorize responses in a way that is required for a traditional confusion matrix, so as in the case of multiple-choice questions, accuracy for also used.
The evaluation of open-ended, synthesis-based questions included the consultation of experts at the UT which participated in the user study described in the corresponding Section 3.5

3.5 User Study

The conducted user study aimed to provide insights into the performance of the LLM on answering open-ended synthesis-based questions.

3.5.1 Methodology. In order to construct the prompt for the aforementioned questions, the Persona Pattern was used [35] in order to give instructions for the LLM on the role it has to adapt and what domain and the course is the question related to. An example prompt format is shown below:

> "Imagine you are a *[the academic year corresponding to the course]* year *[the study programme from which the question was extracted from]* student studying *[the course from which the question was extracted from]* and are presented with the following question at the exam:
> *[the question to be answered]*"

After the needed answers were generated, the participants of the user study were presented with the prompt and the generated output and were asked to grade as they would in a real exam.
Each question presented was graded and analyzed based on two steps. First, the experts graded the question based on three choices:

- Full points (no remarks, the answer every aspect required by the question)
- Partial points (the answer is partially correct, or does not answer the question fully)
- No points given

Following that, the participants were asked to give comments on their answer to the previous question and to specify the good/bad aspects of the generated response that served as basis for their grading.

4 RESULTS

This section presents the outcomes of the evaluation conducted on the performance of LLM in answering exam questions. The responses were evaluated across the aforementioned categories.

4.1 Accuracy across different question formats

According to the results in Table 2 the LLM shows better performance on multiple-choice (MC) than short answer on medium length (71.42% vs 42.85%) and long (42.85% vs 14.28%) questions. The answers to open-ended questions were awarded full points in 50% of the cases and partial points in 37.5% of the cases, which is a rather satisfactory result only failing to provide at least a partially relevant response in only 12.5% of the time.

4.2 Accuracy across different question lengths

The results from Table 2 show varying performance of the LLM on questions of different length and formar with an consistent accuracy only for the short questions (57.14%). Our outputs show that the LLM performed worse on long questions having an average accuracy of 28.57%, comparable to the average of 57.14% for both short and medium length questions.

4.3 Accuracy across different question types

Table 3 indicates a higher average accuracy of 64.28% of factual questions which is higher than on analytical questions with an average result of 42.85%. The results achieved on synthesis-based questions were previously introduced in Section 4.1

5 DISCUSSION

In the following section, there will be a detailed discussion of the results outlined in Section 4. The discussion is structured around the research questions that were initially proposed in Section 1.1 and assesses the implications of the results in the light of the objectives we aimed to address through the research questions

5.1 Research Question 1

Based on the results shown in Table 2 we can derive how the format of the question impacts the accuracy of the LLMs' responses.
The accuracy on medium-length MC questions is the highest (71.42%) potentially showing that this is the optimal balance of the amount of information given. It could also indicate that MC questions are of sufficient length for providing enough detail for effective context comprehension, but without overwhelming the model. Additionally the results show that the accuracy is higher when the prompt consists of factual questions (71.42%) compared to analytical (42.85%) which could indicate that it is more effective for the LLM to retrieve specific information from its training data when the answers should be based on memorizable facts. Additionally, the probability of a model choosing a correct answer even in scenarios where it has limited understanding of the topic is higher than generating an answer on its own.
The performance of the LLM on short answer questions follows a decrease when the question length increases. Multiple factors could lead to such a scenario such as the growing cognitive load on the model when the input question is long as it requires to keep more information in memory and connect all of the parts in order to compute the answer. Additionally, the nature of the short answer questions differs significantly from the MC format, because it requires a concise yet precise, correct answer which could be causing the observed trend. MC question format could be guiding the LLM toward relevant content and significantly reduce the possibility of hallucinations and off-topic responses. Additionally, MC questions would potentially reduce the complexity of the task, as the model would be required to just recognize the correct answer from a set, rather than generate one from scratch. In regards to open-ended questions, the LLM showed the capability to generate a response that was considered complete and ample in terms of detail in 50% of the cases. However, there were instances were the model lacks detail in its response and even speculates. One of the experts from the conducted user study states the following regarding its response for one of the Business Ethics questions:

> "The reference to the company's own code of ethics could be useful, if we would know that the company has this code. Now it is pure speculation."

Additionally, there are cases where the model fails to go beyond the basic interpretations of certain concepts. For example, when discussing ethics dilemmas it associates child labor with the happiness of the families and their benefit from the income:

> "In addition, the Thai workers and their families would also benefit from the steady work and income that the deal would provide. This could result in an overall increase in happiness for a large number of people.**[...]**However, on the other hand, Susan may consider the potential negative consequences of the deal. The use of child labor in the production process goes against ethical standards and could harm the well-being of the children involved. This could also have a negative impact on the company's reputation "

When presented with the task of grading this response, one of the experts from the user study affirms the following:

> "The suggestion of the answer that child labor increases the happiness of the families doing the work is a short sighted, short term interpretation of "happiness". **[...]** For child labor the negative long term effect on children to be excluded from education is much larger than the short term impact of the bonus for Susan."

This depicts a possible drawback of the model when it comes to discussing topics that require deep analysis and interpretation.

5.2 Research Question 2

The results of our research show varying performance of the LLM across different question lengths as per Table 2. Following the discussion from the previous section, our research indicates that that the optimal length of the input question could be dictated by the question format, and it cannot be concluded whether short, medium or long questions are better without taking into consideration the nature of the question given. For example, while the LLM shows a higher accuracy with medium-length MC questions, on short answer questions the LLM tends to perform better on short inputs.

5.3 Research Question 3

When comparing analytical questions to factual, we can observe that the achieved accuracy is lower in both MC and short answer cases (71.42% in MC factual vs 42.85% in MC analytical; 57.14% in short answer factual vs 42.85% in short answer analytical). This trend could be pointing out the added difficulty of processing and interpreting information that requires a deeper understanding or context analysis. One known area that corresponds with this scenario and was represented in the dataset used in this research is the field of mathematics. Research shows [37] that LLMs performance is inconsistent on questions that include varying textual forms that include both words and numbers. Another factor to take into consideration is the length of the factual vs analytical questions. Usually, due to the nature of the factual questions which aim to ask for definitions, facts and specific information, the amount of context needed for specifying the question is much lower compared to analytical questions, where formulas, assumptions and hypothesis have to be described which can overload the model. Based on the results from Fig. 2 the LLM can often handle synthesis-based questions in an effective matter, staying on the relevant topic and providing enough details and explanations to satisfy the examiner's requirements, as it achieved full-points for its responses in 50% of the cases. However, the results indicate inconsistencies in the other half of the cases, when it fails to provide enough detail or does not delve deeper into the question topic. One of the participants of the user study affirms the following about the LLM response on one of the questions regarding the course of Electronic Commerce:

> "Again, half of the response is rephrasing the question. After that the response starts well, however the text ends when examples seem to start."

This could be a suggestion that when the model does not have a deep understanding of the topic, it relies solely on the content of the question to generate an output that would only depict a valid answer at first glance, however would lack depth and detail when analysing it.

> "The response correctly identifies that a Web Servlet (HttpServlet object) uses the Servlet API provided by Java to handle HTTP requests and responses. While the answer is correct in general, it lacks depth in explaining how exactly the HttpServletRequest and HttpServletResponse objects are used in the Servlet lifecycle."

5.4 Limitations

This research aimed to discover patterns in performance of LLMs in exam contexts, however several limitations have to be taken into consideration.

5.4.1 Dataset. First, due to time constrains and absence of direct availability of exams and their solutions from any faculty at the UT, the questions were mainly concerning the Computer Science field for the multiple-choice and short answer type questions and Business Information Technology and International Business Administration for open-ended questions. This makes the results and conclusions limited to these fields and may not be relevant for exam questions from other fields. Additionally, human bias was introduced due to the fact that the dataset was manually constructed, thus potentially leading to a non-representative sample of exam questions.

5.4.2 Evaluation. The expert input used in the evaluation of open-ended synthesis-based question introduces subjectivity as different researchers, professors and teaching assistants might have different standards of what a full and a partially correct answer is. Additionally, the interpretation of the answer might differ for different expects involved. For instance, when grading the same exam question regarding Business Ethics on the topic of business involving child labour one of the participants awarded full points noting that:

> "As there is no right or wrong for the business ethics (as it's personal interpretation), it deserves full points."

While the other examiner awarded no points as they expect the students to be critically aware of the implications of child labor and any response that claims that child labour might be an acceptable practice suggests that the interpretation or analysis of the ethical dilemma is incomplete or needs reconsideration.

Another limitation of the evaluation method is the reliance on accuracy which may not capture partial correctness of the answer in the case of MC questions. Additionally the BLEURT scores may not fully capture the semantic correctness of the generated output and may not account for acceptable variability in phrasing or alternative correct answers, consequently leading to the underestimation of the model's performance in generating valid, however differently phrased responses.

6 CONCLUSION

This research aimed to discover and analyze the patterns of LLMs in question answering exam-style questions. It employed a few-shot methodology with a manually constructed dataset of questions and answers from real examinations at the UT. Through evaluation based on accuracy and expert grading, the study revealed patterns in performance based on different question types, lengths and formats. The findings of this research provide useful insights indicating that the LLM performs best on medium-length multiple choice questions, particularly with factual content, achieving the highest accuracy, suggesting that this format and length balance is the most optimal for the model by providing enough information and context without overwhelming the model. The accuracy is observed to decrease with longer, short answer format of questions indicating a possible weak spot of the model when presented with a higher level of cognitive load while requiring a high precision answer. The study revealed that the model performs on par in 50% of the cases on open-ended, synthesis-based questions, while still generating responses that lack detail, depth of interpretation and include speculations in other 50% of the time, highlighting possible limitations in handling nuanced or ethical topics.

The comparative analysis across different question types depicts a higher level of performance in factual vs analytical questions. This could be explained by the different distinctive nature of these question types, one requiring just retrieving information from the training data and the other heavily loading the model with information and requiring complex deduction of results and analysis.

Ultimately, this study contributes to the identification of strengths and limitations of LLMs in the context of exams. It suggests that future research is needed for exploring techniques for the enhancement of interpretative capabilities of LLMs and provides a foundation for studies regarding improvements of AI in educational applications.

REFERENCES

[1] A. Buscemi and D. Proverbio, "Chatgpt vs gemini vs llama on multilingual sentiment analysis," *ArXiv*, vol. abs/2402.01715, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267412584.

[2] X. Pu, M. Gao, and X. Wan, "Summarization is (almost) dead," *ArXiv*, vol. abs/2309.09558, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262044218.

[3] Y. Tan, D. Min, Y. Li, *et al.*, "Can ChatGPT replace traditional KBQA models? an In-Depth analysis of the question answering performance of the GPT LLM family," in *The Semantic Web – ISWC 2023*, Springer Nature Switzerland, 2023, pp. 348–367.

[4] J. Koetsier. "Gpt-4 beats 90% of lawyers trying to pass the bar." Accessed: 2023-05-03. (Mar. 2023), [Online]. Available: https://www.forbes.com/sites/johnkoetsier/2023/03/14/gpt-4-beats-90-of-lawyers-trying-to-pass-the-bar/.

[5] R. Bhayana. "Chatgpt passes radiology board exam." Accessed: 2023-05-03. (May 2023), [Online]. Available: https://www.rsna.org/news/2023/may/chatgpt-passes-board-exam.

[6] R. Bhayana, S. Krishna, and R. R. Bleakney, "Performance of chatgpt on a radiology board-style examination: Insights into current strengths and limitations," *Radiology*, vol. 307, no. 5, e230582, 2023, PMID: 37191485. doi: 10.1148/radiol.230582. eprint: https://doi.org/10.1148/radiol.230582. [Online]. Available: https://doi.org/10.1148/radiol.230582.

[7] R. Bhayana, R. R. Bleakney, and S. Krishna, "Gpt-4 in radiology: Improvements in advanced reasoning," *Radiology*, vol. 307, no. 5, e230987, 2023, PMID: 37191491. doi: 10.1148/radiol.230987. eprint: https://doi.org/10.1148/radiol.230987. [Online]. Available: https://doi.org/10.1148/radiol.230987.

[8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL].

[9] S. Wang, X. Sun, X. Li, *et al.*, "Gpt-ner: Named entity recognition via large language models," *ArXiv*, vol. abs/2304.10428, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258236561.

[10] F. Wei, R. Keeling, N. Huber-Fliflet, *et al.*, "Empirical study of llm fine-tuning for text classification in legal document review," *2023 IEEE International Conference on Big Data (BigData)*, pp. 2786–2792, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:267144695.

[11] D. Jinensibieke, M. Maimaiti, W. Xiao, Y. Zheng, and X. Wang, "How good are llms at relation extraction under low-resource scenario? comprehensive evaluation," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270560680.

[12] P. Jiang, C. Sonne, W. Li, F. You, and S. You, "Preventing the immense increase in the life-cycle energy and carbon footprints of llm-powered intelligent chatbots," *Engineering*, 2024, issn: 2095-8099. doi: https://doi.org/10.1016/j.eng.2024.04.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809924002315.

[13] L. Huang, W. Yu, W. Ma, *et al.*, *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*, 2023. arXiv: 2311.05232 [cs.CL].

[14] A. Bhattacharjee, R. Moraffah, J. Garland, and H. Liu, "Towards llm-guided causal explainability for black-box text classifiers," 2024.

[15] I. H. Sarker, "LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 40, May 2024.

[16] M. Rosoł, J. S. Gąsior, J. Łaba, K. Korzeniewski, and M. Młyńczak, "Evaluation of the performance of gpt-3.5 and gpt-4 on the medical final examination," *medRxiv*, 2023. doi: 10.1101/2023.06.04.23290939. eprint: https://www.medrxiv.org/content/early/2023/08/16/2023.06.04.23290939.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2023/08/16/2023.06.04.23290939.

[17] I. Chalkidis, *Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark*, 2023. arXiv: 2304.12202 [cs.CL].

[18] P. M. Freitas and L. M. Gomes, "Does chatgpt pass the brazilian bar exam?" In *Progress in Artificial Intelligence*, N. Moniz, Z. Vale, J. Cascalho, C. Silva, and R. Sebastião, Eds., Cham: Springer Nature Switzerland, 2023, pp. 131–141, isbn: 978-3-031-49011-8.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399.

[20] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *ArXiv*, vol. abs/2302.13971, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257219404.

[21] G. T. G. R. Anil, S. Borgeaud, Y. Wu, *et al.*, "Gemini: A family of highly capable multimodal models," *ArXiv*, vol. abs/2312.11805, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266361876.

[22] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[23] J. H. ndez-Orallo, "Ai evaluation: On broken yardsticks and measurement scales," 2020. [Online]. Available: http://josephorallo.webs.upv.es/papers/AAAI_MetaEval_Workshop2020__PAPER-corrected.pdf.

[24] D. Stribling, Y. Xia, M. K. Amer, K. S. Graim, C. J. Mulligan, and R. Renne, "The model student: Gpt-4 performance on graduate biomedical science exams," *Scientific Reports*, vol. 14, p. 5670, 1 May 2024, issn: 2045-2322. doi: 10.1038/s41598-024-55568-7.

[25] A. Madrid-García, Z. Rosales-Rosado, D. Freites-Núñez, *et al.*, "Harnessing chatgpt and gpt-4 for evaluating the rheumatology questions of the spanish access exam to specialized medical training," *Scientific Reports*, vol. 13, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260155442.

[26] A. Taloni, M. Borselli, V. Scarsi, *et al.*, “Comparative performance of humans versus gpt-4.0 and gpt-3.5 in the self-assessment program of american academy of ophthalmology,” *Scientific Reports*, vol. 13, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264589180.

[27] J. Liu, P. Zhou, Y. Hua, *et al.*, “Benchmarking large language models on cmexam - a comprehensive chinese medical exam dataset,” A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Curran Associates, Inc., 2023, pp. 52 430–52 452. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/a48ad12d588c597f4725a8b84af647b5-Paper-Datasets_and_Benchmarks.pdf.

[28] M. Santos and C. Campelo, “Benchmarking quantized llama-based models on the brazilian secondary school exam,” SBIC, May 2023. doi: 10.21528/cbic2023-177. [Online]. Available: http://dx.doi.org/10.21528/CBIC2023-177.

[29] R. Raimondi, N. Tzoumas, T. Salisbury, *et al.*, “Comparative analysis of large language models in the royal college of ophthalmologists fellowship exams,” *Eye*, vol. 37, pp. 3530–3533, 17 Dec. 2023, issn: 0950-222X. doi: 10.1038/s41433-023-02563-3.

[30] K. Singhal, T. Tu, J. Gottweis, *et al.*, “Towards expert-level medical question answering with large language models,” *ArXiv*, vol. abs/2305.09617, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258715226.

[31] OpenAI, *GPT 4*, https://openai.com/index/gpt-4-research/, [Accessed 15-05-2024].

[32] E. Martínez, “Re-evaluating gpt-4’s bar exam performance,” *Artificial Intelligence and Law*, May 2024, issn: 0924-8463. doi: 10.1007/s10506-024-09396-9.

[33] T. Brown, B. Mann, N. Ryder, *et al.*, “Language models are few-shot learners,” H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[34] T. Brown, B. Mann, N. Ryder, *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[35] J. White, Q. Fu, S. Hays, *et al.*, “A prompt pattern catalog to enhance prompt engineering with chatgpt,” *ArXiv*, vol. abs/2302.11382, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257079092.

[36] T. Sellam, D. Das, and A. P. Parikh, *Bleurt: Learning robust metrics for text generation*, 2020. arXiv: 2004.04696 [cs.CL].

[37] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin, *Large language models for mathematical reasoning: Progresses and challenges*, 2024. arXiv: 2402.00157.