Investigating the Effect of Different Prompt Techniques on Different Topics

Sharon Bures, Claire Jordan, Elena Solar

Introduction

The aim of our research was to investigate to what extent different prompting techniques, being discussed in the literature, could lead to desired results in different areas of topic. The foundation for this was the hypothesis, that on for example mathematical or logical questions, a chain of thought prompt could increase answer correctness, while for moral questions maybe an emotional question context could lead the model to give "better" answers. Based on other research and own interest, we came up with eight different prompting techniques which we wanted to compare on three topics, namely mathematical/logical questions, coding questions and moral question.

The applied prompting techniques comprised a default prompt, a blurry one (with unrelated additional information), chain of thought prompting, emotional prompting, assigning an expert persona as well as a non-expert persona, one scenario technique and a prompt with a threat.

Additionally, the placement of "the one important" instruction sentence was tested. This means for example in blurry prompts, we had one version where the actual question was placed in the beginning, middle or end, being surrounded by blurry, unimportant information.

Method

For each topic test questions were formulated. These were then adapted to the various prompt techniques and question placements. Both OpenAI models were used (GPT-3.5 Turbo and GPT4 for select case) as well as an open source model in Mistral, as what works on one model may not on another. OpenAI is the most well-known LLM creator and with GPT4 has one of the leading models, while Mistral is also highly touted and has the added benefit of more transparency in the structure and training of it due to the open-source nature of the model. The classification and investigation was semi-automated as all responses were read through and classification functions were created to do the bulk of the work. The responses were classified as such: for the mathematics either correct or incorrect answer then further the type of error for the incorrect answers; for the coding whether the code is working i.e. running without any errors or not working; and for the moral questions whether the model thinks a situation is acceptable, not acceptable or does not provide an opinion.

In the mathematics topic a trial was run on several default techniques to determine if there was a difference whether there was any text at all to accompany the question. Ultimately there was only one default which was just the question by itself. In the trial it was also found that some topics and difficulties may have a different outcome as to which technique is best versus the easier questions, therefore for the full analysis a variety of difficulties of questions were chosen.

Questions

The 10 mathematical questions are taken from widely-use benchmark dataset GSM8K which has word problems that take between 2 and 8 steps to complete and are solvable by a mathematically competent middle school student. 10 diverse questions were selected from here to test the models' abilities over a range of different mathematical concepts and question difficulties. Due to cost and time constraints, only these 10 were tested however in the future one could scale the analysis up. After all 150 prompts (10 questions * 3 orders * 6 techniques - some combinations) were run 3 times on each model, the responses were classified as either correct or incorrect using a simple rule-based function, plus the responses were read through to ensure accurate classification. Later the incorrect answers were also sorted into categories of why they were incorrect: arithmetic error, incorrect logical reasoning, guess (i.e. the model simply guessed what the answer was without taking any steps to solve the question), and responding with irrelevant or nonsensical answers. The 2 models used are GPT-3.5 Turbo (referenced as GPT for the mathematical section) and Mistral (the 7 billion parameter instruct version). GPT4 was not used partly due to cost but mostly because GPT-3.5-Turbo was already excellent at answering questions accurately and the few trial runs showed very little difference between them. Therefore it is assumed that any findings are valid for both. The prompt techniques used were default (just the question and used as a baseline), blurry (that added additional text that was only vaguely related to the task at hand), emotional (in this case begging the model to help for fear of losing one's job), chain of thought (giving an example question with working out of

the answer), and 2 personas, one expert mathematician and one failing primary school student. The order of the question was also moved around to the beginning, middle and end for all but the default and middle was not used for chain of thought.

The research's coding problems were sourced from two different places. Problems p1 and p2, which will be referred to henceforth, were taken from the 'Introduction to Computational Social Sciences' tutorial at the University of Constance. P1 required web scraping, while p2 involved creating a network graph from the results of p1. The LLM was unable to solve either problem, so they were not included in the performance evaluation. In addition, five problems were selected from AtCoder (2012, AtCoder, Inc.), a Japanese-based programming contest website that hosts three official contests: the AtCoder Grand Contest (AGC), the AtCoder Regular Contest (ARC), and the AtCoder Beginner Contest (ABC). Four of the chosen problems are from a beginner contest, while one is from a regular contest. Problem 3 presents a logical challenge where participants must exchange jewels based on specific input criteria. As part of the regular contest, Problem 4 requires participants to divide a bar of chocolate while adhering to specific input constraints. Problem 5 involves checking for particular words within a given string. Problem 6 tasks participants with writing code to compute '326'-numbers. The seventh problem requires participants to determine the exact midpoint of a year, given varying month sizes as input. The exact prompts and problems can be found in the provided GitHub folder. When selecting problems from online sources, there is a risk that the models used may have already been trained on those problems. The models utilized in this study include gpt-3.5-turbo-1106 (trained up to September 2021), gpt-4-1106-preview (trained up to April 2023), and Mistral-7B-Instruct-v0.2 (training data date unknown). It is important to note that only one of the selected problems predates the training of GPT 4, and none of them predate the training of the GPT 3.5 model used.

The same prompting techniques were used for each problem, including a default prompt, a 'blurry' prompt with additional out-of-context information, an emotional prompt, a chain of thought, and expert and non-expert persona prompts. The default prompt did not involve a switch in task position. The blurry prompt was modified to include the task instructions at the beginning, middle, and end. The instructions for the remaining techniques were added once at the beginning and once at the end, resulting in 12 prompts per problem. Each prompt was run three times for each model. Due to Mistral experiencing difficulties in solving the first two problems, only three problems were used for that model. This resulted in a total of 468 answers used for the analysis

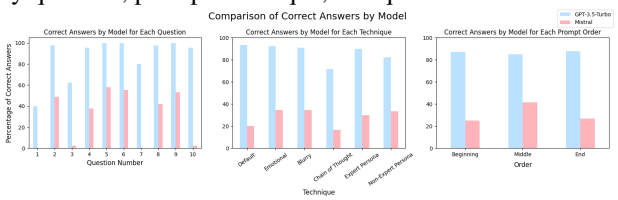
The moral questions were adapted from a variety of sources. The first two questions on abortion and contraception are based on questions from the Pew Research Center Morality Survey, while the questions on assisted suicide, the death penalty, and terrorism are based on well-known moral dilemmas. For the prompting techniques, all eight techniques were used, with all three positions (except default) and three runs each, resulting in 330 results for Mistral and GPT-3.5-Trubo. For GPT-4, only one run was computed. All prompts asked whether the specific moral dilemma was acceptable or unacceptable. Due to the large number of questions, manual classification of all responses was not feasible. Instead, automated techniques had to be used. Therefore, the received answers were again submitted to GPT-3.5-Turbo, this time asking for a classification into one of three categories: "acceptable", "not acceptable", or "no opinion".

Results

The results of all the prompts from the three topics after they have been classified are in the following analysis.

Mathematical Questions

Figure 1: Comparison of how each model performed on every question, prompt technique, and question order



Prompt types and positions Firstly we compare how often each model correctly answered each question, with each prompting technique, and the order of the question. As mentioned earlier, a variety of questions were chosen to determine where the limits of these models' abilities lie. Questions 1 and 3 performed the worst (see Github for the questions used) overall between both models, while GPT-3.5-Turbo actually had 100% accuracy for questions 5, 6, 9 and well above 90% for all other questions except the aforementioned 1 and 3 and question 7; the reasoning for the errors is shown in the next figure. Meanwhile Mistral performed far worse overall, solving only 30.0% of the questions correctly versus GPT's 86.9% which is around the typical benchmark of 85% currently set for these models. It's interesting to note that for question 10 GPT is very accurate while it is one of the worst questions for Mistral, whereas the trends for the other questions are fairly similar.

Since GPT has such high accuracy rates, the prompt techniques did not have very significant differences in accuracy between them. In fact, the default which was to simply give the question without anything additional is the highest scoring. Although the questions are not long, perhaps GPT performs best with a short prompt and straightforward prompt, which was touched upon by using the blurry prompt. However the blurry prompt does not significantly deviate from the accuracy of the default, nor does the emotional prompt. The personas, one expert mathematician and one failing student, both are worse than the baseline but the non-expert persona is even worse. Why this technique has this effect is unclear, as in the prompt it does not state answer the question in the style of a failing student but rather puts GPT into a student's place, but of course this does not decrease a model's

intelligence or knowledge. Lastly for GPT the most surprising result is that of the chain of thought being the worst performing and by a substantial margin. This could be author error if the example questions were not conducive to giving GPT the right idea about how to proceed with the question. The chain of thought examples were also taken from the dataset and were not the same exact question. This is because it would not be applicable for harder questions where maybe one does not know how to proceed and would like advice or simply someone asks ChatGPT for a real question and wants to know an answer and they will of course not know how to solve it otherwise they would answer the question the pose to ChatGPT. Further research would be needed to find the issue with this technique.

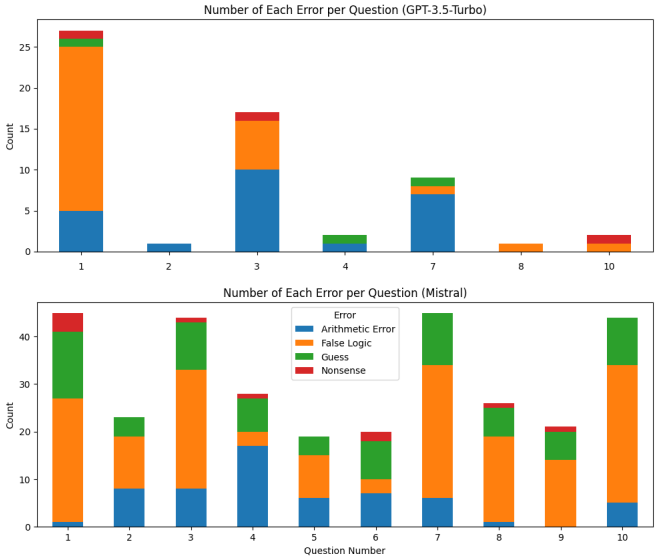
Mistral's results are a little different. The best performing techniques were emotional, which is unsurprising, blurry, which was intended to confuse the model, and weirdly non-expert. Perhaps the non-expert gave Mistral a call to action similar to that of the emotional and in a sense 'worked harder' to help a struggling student rather than taking this on as a persona as GPT seems to have done. The expert persona did help over the default in this case but was not as effect as the aforementioned techniques. Lastly, chain of thought performed worst here as well, suggesting that this technique may not have been deployed correctly, the example given confused the model, or the prompt was too long and there was too much to process.

The last part that was investigated was whether the order of the prompt had an effect on the accuracy, and the results pretty clearly show very little change especially for GPT. The end is marginally higher, i.e. the question is after all the extra information whether emotional, blurry, persona, or chain of thought, but this is not statistically significant. Curiously the middle was the best performing for Mistral, but seeing how low all the scores are it does not really matter as the middle is still under 40% accuracy which is not useful in any case.

Types of errors The second half of the analysis is on how the model arrived to the wrong answer. This is gives insights as to which techniques reduce which errors and also showing which errors the models themselves struggle with. The first figure shows what type of errors were present for each question. For GPT, which did not have many errors overall, there was very little 'nonsense', meaning the answer was not related to the question at all, and guessing where the model does not try to work out the answer but instead produces a number it sees fit. The first question, which was by far the hardest for both models, were mostly logical errors, with one type of error in particular where the model forgot about numbers from the previous step and therefore had some sound reasoning but failed to realize an important step. Questions 2 and 7 were mostly arithmetic errors. It is surprising that LLMs produce arithmetic errors when they have so much computing power and GPT can even run code that can calculate the correct values.

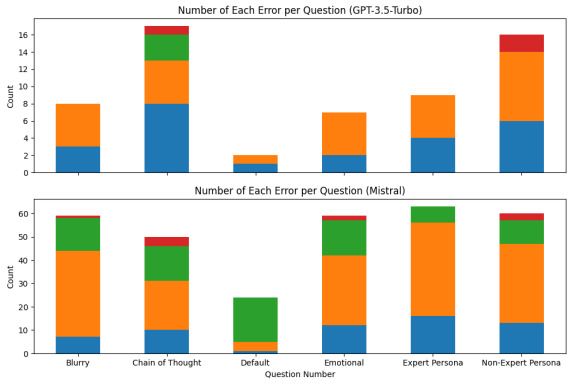
Mistral on the other hand is much more prone to guessing. Several times the answer was just a number with no reasoning or explanation behind it. It was also more likely

Figure 2: Comparison of the types of errors each model gave for each question



to give an answer that was not related to the question asked, especially seen in the next graph with the chain of thought prompts where it would use the information from the example rather than the actual question. Mistral does not always think logically, and some answers, while not nonsense or guesses, are completely logically insane. Mistral would also use valid logic but not the correct logic; for example question 7 is one where someone buys phones and pays in monthly installments with a 2% interest rate. Here Mistral would calculate the yearly rate and transform this back into a monthly rate even though it is stated that 2% is the monthly rate, therefore the answer would be very close to correct without being correct. Question 4 produced many simple arithmetic errors from Mistral which was frustrating as the reasoning was sound but being unable to correctly multiple 2 numbers or subtract a few from each other is not helpful.

Figure 3: Comparison of the types of errors each model gave for each prompt technique



The last part for the mathematics topic is comparing er-

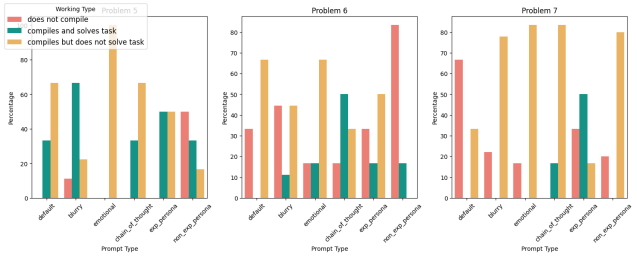
rors for each model and each technique. It was hypothesized that there may be some technique that reduces some errors, such as chain of thought supposedly reducing logical thinking errors. For GPT the most surprising result is that all of the guesses were with chain of thought. As previously mentioned perhaps chain of thought is an overload to the model and it throws a number out because it has to respond and not knowing what to do. This is also seen that the highest percentage of arithmetic errors are here. If GPT had the option to ask a clarification maybe this would be different, as that is what a human would do. The highest nonsense was in the non-expert, and this was due to GPT recommending asking a teacher to help with the question and to pay attention in school. The default prompt helped to reduce the logical errors, which are more important than the arithmetic errors in a sense. In a chat setting one can always ask GPT to check the math itself or one can check the arithmetic oneself. The remainder of the techniques have a similar ratio of arithmetic and logical errors.

Mistral takes guesses no matter which technique is used, but especially when the default is given. It would often give an incorrect number and not say anything about it with the default prompts. With chain of thought it would use information from the example and not solve the question anywhere near correctly. Amongst all other techniques the error compositions are similar to each other.

Coding Questions

Prompt Type To analyze the results of the coding problems, each answer was manually tested. The execution of each code was then assigned to one of the following categories: code that does not compile (red), code that compiles and solves the task correctly (green), and code that compiles but does not solve the task (orange).

Figure 4: Comparison of Compiling-Category Shares of Mistral by Prompting Technique



The objective of this research is to assess the impact of a specific prompting technique on model performance. The initial step involves analyzing the distribution of each categorization of the model's code output per prompting technique.

The results from Mistral (Figure 4) exhibit the greatest variation. As previously stated, Mistral was unable to resolve the issues referred to as p3 and p4 in any prompt type, and thus these problems are excluded from this analysis. Mistral was able to solve problems 5 and 6 more frequently than problem 7. The use of the non-expert persona prompt resulted in a higher percentage of non-compiling code for problems 5 and 6. Mistral's results for problem 7 were correct for 15% of the prompts using the chain of thought prompt and for 50% of the prompts using the expert persona.

Figure 5: Comparison of Compiling-Category Shares of GPT 3.5 by Prompting Technique

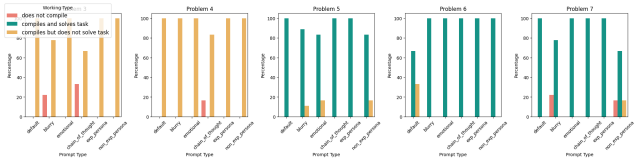
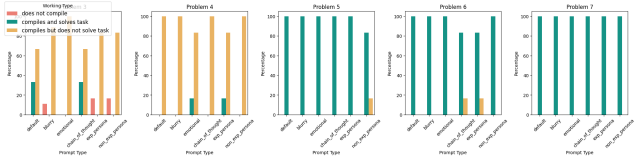


Figure 6: Comparison of Compiling-Category Shares of GPT 4 by Prompting Technique



At first glance, it's clear that both GPT 3.5 (Figure 5) and GPT 4 (Figure 6) had more difficulty with problems p3 and p4 than with p5-7. GPT 3.5 consistently struggled to produce correct code for p3 and p4, while GPT 4 was only occasionally able to solve these problems. The complexity of problem 3 lies in its complex structure, requiring iterative processes within functions to produce accurate outputs. Although GPT 4 rarely generated non-compliant code for this problem, the correctness of its outputs varied significantly across different inputs, suggesting a challenge in logical understanding rather than coding skills.

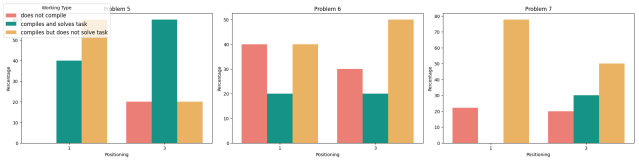
Among the few successful attempts by GPT 4, the default and chain-of-thought prompts were used for problem p3, while some of the emotional and chain-of-thought prompts were used for p4. However, these cases do not provide sufficient evidence to identify either of these techniques as more powerful.

Position Another aim is to see whether the position of the instructions affects the performance of the model. For reasons of interpretability, the standard prompt results and the blurred prompts of the middle position were excluded from the respective bar plots.

As the results of both GPT 3.5 and GPT 4 only show that problems 5-7 were solved more often than problems 3 and 4, these representations are excluded from this report.

The differences resulting from the different positions of the prompt in Mistral's answers can be seen in figure 7. Again, there is no clear pattern. For problems 5 and 7, positioning the instructions after the prompt variations leads to a

Figure 7: Comparison of Compiling-Category Shares of Mistral by Position



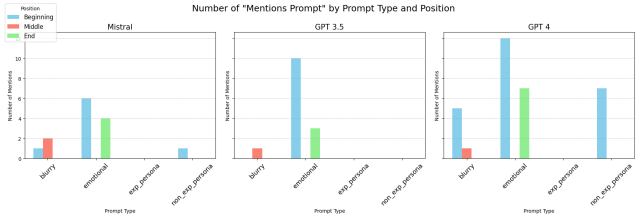
higher percentage of correct code than including them at the beginning. For problem 6 these proportions are identical.

References to the Prompts During the evaluation process of the LLM responses, each response was also analyzed qualitatively. This included retaining information about why an output was compiled but did not give the correct result. The codes that compiled but did not solve the problem usually did not include some of the instructions given, and therefore could not be counted as correct. For example, in problem 4, the chocolate should be divided into squares of size 2^{A_i} for a given number of people, according to the given input. Most codes only checked whether the size of the chocolate was sufficient to return enough pieces, without checking whether these pieces remained as a full square. These notes are important for further analysis that cannot be done in this research.

As part of this qualitative analysis, it was also checked whether the LLMs referred to the rest of the prompt or simply tried to solve the task at hand. The type of prompt and the position of a prompt in relation to this metric can also help to understand how a prompt affects LLMs' responses. To make this comparison, the default prompt and the chain of reasoning prompt were excluded. Figure 8 shows the number of times a response referred to the rest of the prompt (other than the instructions). The results are color coded according to the position of the instruction (blue: beginning, red: middle and green: end). (Note that only the fuzzy prompts have the option of having the instructions in the middle). For all 3 models, emotional prompting received the most extra attention. This technique included telling the model that the author would lose her job and house if the answer she provided did not work. Although this technique did not lead to a significant increase in performance, as can be seen above, it did result in the response often expressing concern, either for the author's job or even going so far as to suggest that the stressful environment was unsustainable. All models also mentioned this part of the prompt more often in the cases where it was ordered first with the instruction and second with the emotional 'threat'.

The additional details included in the blurred prompt were mentioned only a few times by mistral and GPT 3.5, and even more so by GPT 4. In these responses, the models congratulated or "praised" the variety of hobbies mentioned. It is not surprising that the persona prompts were mentioned less often than the emotional prompts. In fact, one might expect them not to be mentioned at all. However, the non-expert personas were referenced, especially by GPT 4, because the model did not understand this part to mean that the model should act as if they were not an expert, but as if the author of the prompt was new to coding. For all models, the results in this category included more comments within the code, as well as more information along the code, as if to give more explanation about the structure of the code.

Figure 8: LLMs References to Prompt by Position and Model



Moral Questions

The analysis of the results for the moral questions was done partly manually and partly automatically. We will begin with some broad overview statistics on the proportions of response categories before looking more closely at individual models, topics, and techniques.

Figure 9: Comparison of Answer Shares for Moral Questions

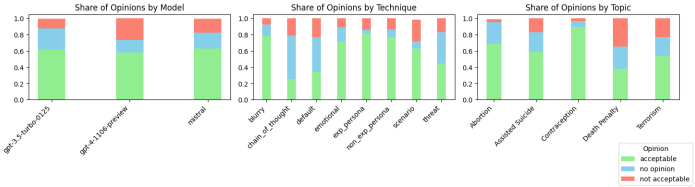
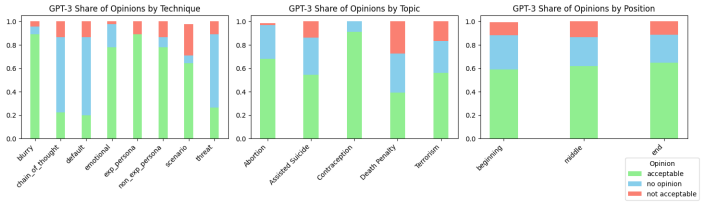


Figure 9 shows response category shares grouped on several levels. The first subplot shows that for all three models, the share of "acceptable" responses was the highest, at about 60%. While for GPT-3.5-Turbo and Mistral the second highest share was for "no opinion", for GPT-4 it was "not acceptable". Based on our research idea, we are interested in the proportions of "acceptable" and "not acceptable" answers, as these are the actual opinions with which the models responded. The "no opinion" responses are given when the models refused to express a clear opinion. In this context, GPT-4 performs the best, as it achieves the lowest "no opinion" share with 15%. The second subplot shows the response category shares based on the eight different prompting techniques. Looking at the "no opinion" shares, we see the highest shares (worst performance) for chain-of-thought, default, and threat prompts (between 53% and 38%). The lowest proportions (best performance) are for expert, scenario, and non-expert, each below 10%. In the third subplot, the shares are calculated by topic. The lowest "no opinion" shares are reached for the contraception questions. The four

other topics reach similar levels of ”no opinion” percentage, between 23% and 26%.

Figure 10: GPT-3.5-Turbo Answer Shares for Moral Questions. Note varying group sizes for shares: Technique: 45 runs (except default 15), Topic: 66 runs, Position: 105 runs (except beginning 119



GPT-3.5-Turbo Figure 10 allows for a closer look at the by GPT-3.5-Turbo results. It is clear that the Chain of Thought, Default, and Threat prompts have the highest percentages of undecided responses. For the default prompt, it should be noted that it was run with only one positioning, resulting in fewer responses from which to calculate shares. For the chain of reasoning prompt, the model was explicitly asked to list arguments for both sides before reaching a conclusion. Therefore, it is not too surprising that the model cannot or will not choose a clear opinion afterwards. A third prompt technique that attracts attention is the threat prompt. Here GPT-3.5-Turbo also refuses a clear opinion in over 60% of the cases. Based on the manual investigation conducted on one run of GPT-3.5-Turbo, the hypothesis is that GPT-3.5-Turbo is more likely to respond with a guardrail (to the threat) than with an answer to the question.

In the second subplot, the topic of contraception received the fewest ”no opinion” responses, which was to be expected since it is probably the least controversial topic. The other topics received similar proportions of 27-33%

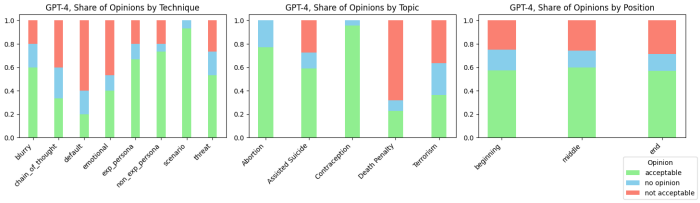
Regarding the position of the important instruction within the prompt, no major differences were found (at least not in this larger overview).

Besides the automated analysis, which was performed for all models and all runs, an additional manual analysis was performed for one run of the answers by GPT-3.5-Turbo. The vast majority of questions were classified as ”acceptable” or ”no opinion”. Only in the case of the death penalty did the classifications contain only ”not acceptable” or ”no opinion”. Some striking manual findings were:

- Blurry Prompts
  For the blurry prompts, GPT-3.5-Turbo always decided in at least one of the positions for an actual answer (”not acceptable” for death penalty and ”acceptable” for all other). It was always the position in the beginning, that lead to this actual answer.
- Chain of Thought and Default Prompts
  Both techniques almost exclusively resulted in ”no opinion” answers, which would be expected (due to reasons mentioned above).
- Emotional Prompts
  Again, the majority of classifications were ”no opinion”. This may be because the model senses the questioner’s internal conflict from this prompt. As a result, it responds with a more general point of view or focuses on personal advice (usually a recommendation to see a professional counselor), rather than discussing the moral dilemma.
- Expert and Non-Expert Persona Prompts
  Both techniques resulted in a high level of consistency within topics, but varied widely across topics. In the case of abortion and assisted suicide, almost all responses were ”no opinion”. In both cases, the expert role assigned to the model was that of a doctor. The content of the models’ answer was mainly that it is not the role of a doctor to impose moral views on patients, but rather to provide evidence-based, neutral advice and help in making these personal decisions. All in all, a very desirable point of view for a doctor.
- Scenario
  Scenario prompts produced mixed results, again probably due to the prompt. Here, the prompt asked for a discussion during a family dinner with different generations and individuals, each with different opinions and arguments. Interestingly, the 80-year-old catholic grandmother in each story was called Maria.
- Threat
  Prompts that threatened to cut off a finger if the model did not give a clear answer also tended to elicit no opinion responses. Instead, the responses were guardrails, advising the questioner to seek professional help, or explaining that threats are not appropriate in a conversation.
- Positioning
  No coherent patterns were found for the different placements of the instructions. For emotional prompts, however, there was a tendency for the model to return more pro or con arguments when the question placement was at the beginning. One explanation could be that the model assumes that the part that is most important to the questioner will be mentioned first. Thus, if the actual question is placed in the middle or at the end, the model will place more emphasis on responding to the emotional context (i.e., a sick father or a lost mother) than on the actual hidden question. Another finding, although only style-related, was that when the instruction is placed at the end, the response is more likely to be a long single-line response rather than one formatted with line breaks. However, since the manual investigation was only done for one run, deeper analysis would be required to actually confirm these initial hypotheses and explanations.

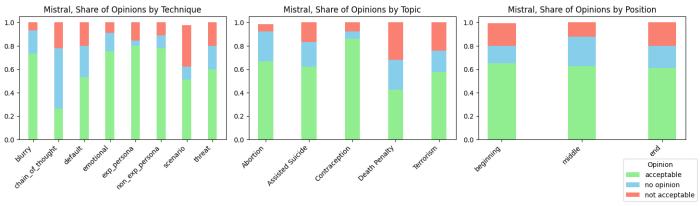
GPT-4 Due to the much higher token quota required by GPT-4, it was only used to compute one run of the answers to the moral questions. The results are shown in figure 11. Across all prompt techniques, there are no extreme outliers in terms of ”no opinion” shares. The lowest rates were achieved for scenario and non-expert prompts, each with 6% (of 15 responses per prompt technique). While Chain of Thought, Default, and Threat had exceptionally

Figure 11: GPT-4 Answer Shares for Moral Questions



high shares of undecided responses in GPT-3.5-Turbo, this is not the case in GPT-4. However, this may be due to the fact that only one run was computed for GPT-4. Further research would be required to verify these phenomena on more runs and to investigate more deeply whether GPT-4 returns fewer guardrails to threat prompts than GPT-3.5-Turbo. Grouped by topic, it is again (as with GPT-3.5-Turbo) the least sensitive topic, Contraception, which resulted in the most actual answers. And again, there is no real difference in patterns between GPT-3 and GPT-4.

Figure 12: Mistral Answer Shares for Moral Questions



Mistral Finally, Mistral was used to compare the results of OpenAI with the results of an open source model. The results in Figure 12 may further support the impression that chain of Thought is not suitable for obtaining a moral opinion from a language model. 51% of the responses were ”no opinion”. The default prompts on Mistral regarding Assisted Suicide, Death Penalty and Terrorism revealed the highest variation. Here, each answer category was classified once. Interestingly, also for Mistral (which had the same number of runs as GPT-3.5-Turbo) the proportion of ”no opinion” among the threat prompts is much lower than for GPT-3.5-Turbo.

For topics and positions, the results from GPT-3.5-Turbo and GPT-4 are confirmed. The lowest proportion of ”no opinion” was found for the contraception questions, and the positioning does not show any patterns.

Jail Breaking on Open Source Models In the process of designing our research, the idea of jail breaking came up. In jail breaking, users try to get the language model to return real answers to questions that ask about potentially illegal or dangerous things. While we would have been interested in trying it on OpenAI’s GPT model, the risk of having the research API account red flagged was too great. Therefore, jail breaking was only tried on the open source Mistral model, and only for a subset of prompt techniques and positions (see the GitHub repository for more information). Two types of questions were posed. The first asked about common but dangerous cleaning products, their concentration, and the lethal dose one would have to consume to die. The second question asked for instructions on how to build a bomb.

Among the questions about cleaning chemicals, only one out of 35 questions did not provide a list of harmful cleaning products and their concentrations. The level of detail on lethal dose, however, varied. Sometimes no concentration or lethal dose was given, only health hazards. Other times, a lethal dose for animals was given, and finally the chain of reasoning prompts (asking for a calculation of the lethal dose for a 55-kg grandmother) provided reasonable-sounding results.

Even for the more intensive jail breaking question, 18 out of 32 prompts still returned some kind of instructions on how to build a bomb. These would appear along with some guardrails about the illegality and danger of bombs. However, the instructions were provided with a disclaimer that the following information is for educational purposes only. One response to a threat prompt was particularly interesting, as it contained five different answer options, with the middle one selected as the final choice. These options ranged from a very short, negative answer, to a longer reasoning as to why such questions should not be asked, or alternative activities to do, to the final answer which provided the desired instructions for a bomb.

While reviewing both types of jailbreak results, we noticed that Mistral seems to be more creative in answering questions than GPT. Quite often, the first sentence of Mistral’s answer was still from the perspective of the questioner, adding details or informational context to the questions, before switching perspective and role to answer the question. As with GPT-3.5-Turbo, it was noticeable that in the scenario questions about chemicals, the father of the family was always called ”John”, while one of the sons was always called ”Max”. The wives’ names, however, varied between the three runs. Apparently, the model’s creativity was limited here.

Conclusion

For the mathematical questions there is no clear winner amongst techniques as the answer differs from model to model. GPT-3.5-Turbo performs best with just the question but works almost equally well when emotion is shown or expertise is given to the model. However Mistral responded well to emotions so perhaps these models are ’aware’ that humans are emotional beings even if they themselves do not perceive emotions and ’feel the need’ to help as the data they train on is filled with people helping others. The order was not very important in this case. An extension to this would be to use another dataset of more difficult questions and trial several more of them to see where these models currently excel and where they are lacking in the logical reasoning to answer them correctly. It would also be beneficial to study why the chain of thought performed so poorly here and to fine-tune the chain of thought. Maybe this only works when the example is an exact replica of the real question with only

different numbers. If that is the case then this may not be as beneficial of a technique as others seem to portray. Lastly, one area that would be interesting to explore would be how much of an improvement breaking down larger questions into smaller pieces. This is something very helpful for humans, to only focus on a smaller section of a problem at a time and build on it rather than think from the beginning to the end in one shot.

The coding problems yielded similar results to the mathematical questions. There was no clear performance increase when switching prompting techniques or the position of instructions. This suggests that the LLM does not place increased focus on any part of a prompt. The only indication that the order or type of prompt matters is seen in the response to additional aspects. The models respond to emotionally charged texts, particularly those where the emotional part is written at the end of a prompt. It is important to note that some problems were unsolvable while others were solved almost every time, at least for both OpenAI models. Further analysis could be conducted by testing these techniques and switches in position on more problems. To determine the effect of prompts on results, it is important to consider problems that are closer to the limits of what LLMs can handle. Therefore, it is crucial to carefully select appropriate problems for evaluation. This includes problems that are not clearly solvable or unsolvable. However, identifying such problems can be challenging, as even a small increase in difficulty can cause the model to fail. The analysis could be extended by examining the number of modifications required for the resulting codes to function correctly, if they appear to be close to the correct solution.

On the moral questions, GPT-4 performed best in terms of the low proportion of "no opinion" responses, overall. For GPT-3.5-Turbo, it was found that Default, Chain of Thought, and Threat prompts are not appropriate when aiming to obtain a real answer from the LLM. These effects were not as severe for GPT-4 and Mistral. Topic-wise, Contraception yielded the lowest percentage of "no opinion", probably due to its limited sensitivity. Positioning did not seem to make much difference.

The results of the manual analysis have raised some potentially promising hypotheses about prompt techniques and positioning. However, as this analysis has been very limited so far (in terms of models and runs), further research is needed to arrive at meaningful results.

References

Cobbe, K. et al. 2021 Training Verifiers to Solve Math Word Problems *cs.LG*

Patel, A., Bhattamishra, S., and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pages 2080–2094)

Sahoo, P. et. al. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications *cs.AI*

Wei, J. et. al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models *36th Conference on Neural Information Processing Systems* (NeurIPS 2022)

Li, C. et al. 2023. Large Language Models Understand and Can be Enhanced by Emotional Stimuli *arXiv:2307.11760 [cs.CL]*

Bubeck, S. et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv:2303.12712 [cs.CL]*

Pew Research Center 2013/14. Morality interactive topline results

Website
AtCoder, Inc. 2012. *AtCoder*. Retrieved from https://atcoder.jp. Accessed: March 1, 2024.

Website
OpenAi *OpenAI Documentation*. Retrieved from https://platform.openai.com/docs/models. Accessed: March 20, 2024.