Are Large Language Models Still Distractible?

Abstract

Large Language Models (LLMs) are increasingly used for complex reasoning tasks, yet their vulnerability to irrelevant context, also known as distractors, remains underexplored. This study investigates the impact of distractors by constructing an adversarial dataset, GSM1C, derived from GSM8K. We evaluate GPT-3.5 and GPT-4o across multiple reasoning steps and prompting strategies, revealing GPT-3.5's limitations and GPT-4o's relative robustness. These findings highlight the potential shortcommings of LLMs as natural human language is prone to containing irrelevant context.

1 Introduction

Large Language Models (LLMs) like GPT-3.5 and GPT-4 have shown remarkable capabilities in reasoning tasks (Shi et al., 2023). However, their performance under conditions with irrelevant context, such as distractors, is not well-understood. This study addresses the following questions:

- How do distractors affect the reasoning performance of LLMs?
- Can LLMs explicitly identify and ignore distractors?
- How do various prompting strategies mitigate the effect of distractors?

To answer these questions, we create an adversarial dataset (GSM1C), systematically evaluate the two LLMs under varying conditions, and analyze their performance with different prompting strategies.

2 Dataset Construction

We modified the GSM8K (OpenAI, n.d.) dataset to create GSM1C, introducing eight types of distractors:

- **Adding Operation**: Introduced irrelevant calculations.
- **Distractor Insertion**: Added plausible but irrelevant sentences.
- **Number Substitution**: Replaced original numbers with misleading values.
- **Integer-to-Fraction Conversion**: Altered number formats.
- **Digit Expansion**: Increased numerical values.
- **Reversing Operations**: Rephrased questions to change reasoning structure.

Example (Adding Operation):

> **Original Question:** Jessica is six years older than Claire. In two years, Claire will be 20 years old. How old is Jessica now?
> **Modified Question:** Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, Claire's father's age was three times Jessica's age.* How old is Jessica now?

3 Evaluation Methods

3.1 Models Evaluated

We tested GPT-3.5 Turbo and GPT-4o, focusing on their reasoning accuracy under original and distractor-injected conditions.

3.2 Metrics and Prompting Strategies

Metrics:

- **Accuracy**: Percentage of correct answers across datasets.
- **Error Analysis**: Confusion matrices for distractor detection.

Prompting Strategies:

- **Chain-of-Thought (CoT)**: Step-by-step reasoning prompts.

- **Zero-shot CoT**: Relied on inherent model reasoning.

4 Results and Analysis

4.1 Performance by Prompting Strategy

Figures 1 and 2 illustrate how prompting strategies perform for GPT-3.5 and GPT-4o across different step counts. GPT-3.5 exhibits significant performance drops when distractors are introduced (GSM1C dataset), while GPT-4o maintains consistently high accuracy across all conditions. These findings emphasize GPT-4o's superior reasoning capabilities and its robustness against irrelevant context.

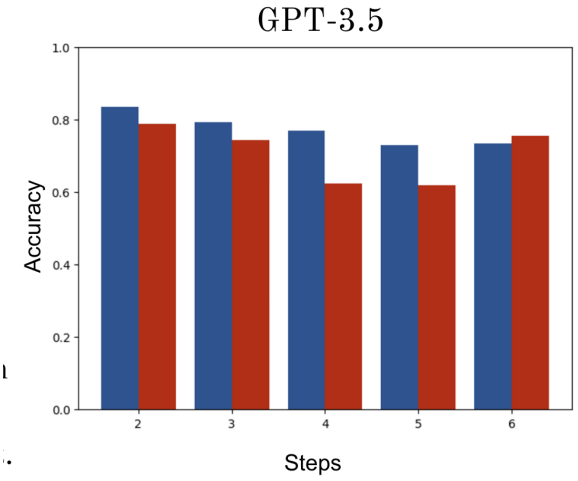

Figure 1: Accuracy of GPT-3.5 across step counts for GSM8K and GSM1C datasets. Accuracy drops significantly for GSM1C, reflecting GPT-3.5's sensitivity to distractors.

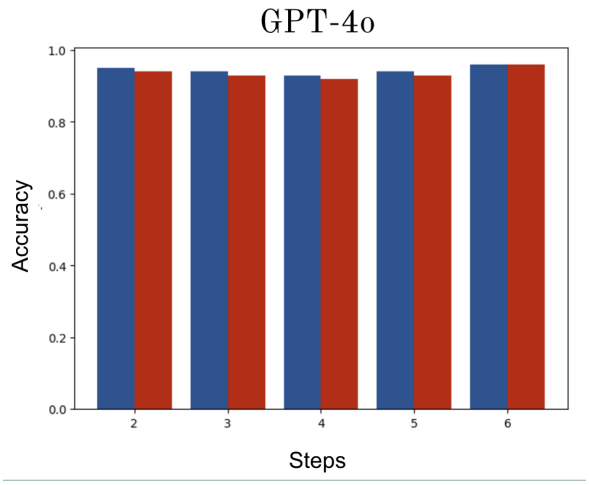

Figure 2: Accuracy of GPT-4o across step counts for GSM8K and GSM1C datasets. The model demonstrates consistent and robust performance across all conditions.

Neither we nor the GSM8K paper (Shi et al., 2023) can offer a definitive reason why the models we tested tend to actually increase in performance when the problems contain 5-6 algebraic steps, especially for GPT-3.5. One would think accuracy would decrease as the number of steps increase, which does happen from 2-4 steps on the models we tested. We manually explored the outputs for both models and didn't find any glitches or red flags in the inputs/outputs. We did find that between the two models, their average responses were about 7% longer over our GSM1C dataset than on the original GSM8K dataset: 304 characters vs. 283 characters per response. This perhaps suggests that our distracting sentences encourage the models to explain their reasoning further than without the distactions - although the overall weaker performance suggests on our GSM1C dataset suggests this isn't always good.

4.2 Investigation of Errors

Figure 3 analyzes model performance under different distractor configurations, focusing on Role, Number, and Sentence Topic. For GPT-3.5, role overlap and out-of-range numerical values significantly reduced accuracy, reflecting its susceptibility to irrelevant context. GPT-4o, however, demonstrated greater robustness, maintaining higher accuracy across all distractor types.

Notably, both models struggled with in-topic distractors, indicating a shared difficulty in filtering contextually plausible but irrelevant information. These findings highlight the need for advanced reasoning techniques to mitigate the impact of distractors.

4.3 Can It Detect Distractors?

Figure 4 evaluates GPT-3.5's ability to detect distractors in input questions. While the model demonstrates strong performance in recognizing non-distractor contexts (2479 true negatives), it struggles significantly with distractor detection, as evidenced by 2351 false negatives. This highlights the model's reliance on surface patterns and its difficulty in explicitly addressing distractor influence.

4.4 Performance Across Steps

Figure 5 illustrates the accuracy of different models as the number of reasoning steps increases, based on both the original dataset (GSM8K) and the distractor-injected dataset (GSM1C). GPT-3.5 Turbo exhibits a steady decline in accuracy as steps

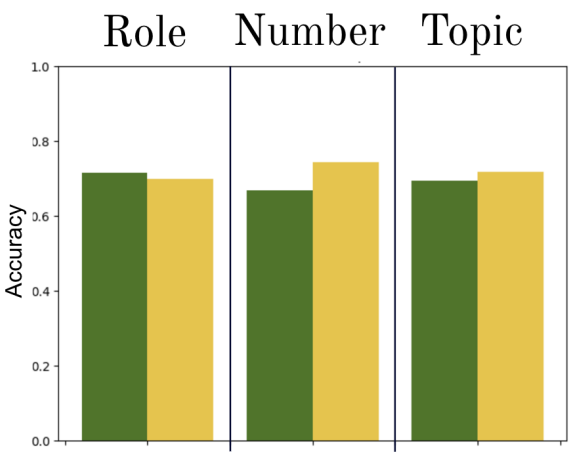

Figure 3: Accuracy across Role, Number, and Sentence Topic distractors for GSM8K (original) and GSM1C (distractor-injected). GPT-4o shows greater robustness to distractors compared to GPT-3.5.

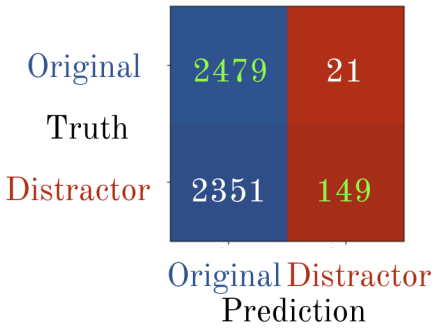

Figure 4: Confusion matrix for GPT-3.5's distractor detection performance. The model exhibits a high false negative rate, failing to identify distractors in most cases.

increase, reflecting its difficulty with multi-step reasoning. The slight improvement at 6 steps is likely due to repetitive patterns in the dataset that enable memorization rather than genuine reasoning.

In contrast, GPT-4-Mini and GPT-4o maintain consistently high accuracy across all steps, highlighting their superior contextual understanding and reasoning capabilities. The inclusion of distractors in GSM1C results in lower accuracy across all models, underscoring the challenge posed by irrelevant context.

We ran an additional experiment on GPT-3.5 with the distractor-injected dataset where we included in our prompt that a distractor was present. Interestingly enough, by simply doing so, we saw an increase in model accuracy (Figure 5).

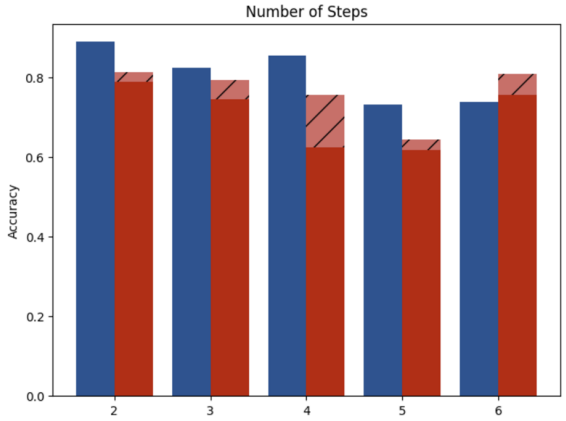

Figure 5: Accuracy trends of different models as reasoning steps increase. Models evaluated on GSM1C (distractor-injected) demonstrate reduced performance compared to GSM8K (original). The striped red bars indicate the performance increase by indicating the presence of a distractor in the prompt.

4.5 Comparison of Original and Modified Sentences

Figure 6 compares the attention maps between the original and modified sentences. Cross-attention between the output tokens and input tokens is shown. The average across all layers and heads is used.

We can see that the attention maps are quite similar, with the exception of a few columns in the middle. It turns out that these are exactly where the output tokens attend to the input tokens of the irrelevant sentence, and the attention scores are very low. This reveals that the output tokens in general do not attend much to (i.e. are not distracted by) the irrelevant input tokens.

Surprisingly, this turns out even to be true for questions that the model answers incorrectly (Figure 7), which refutes the hypothesis that the model is attending too much to irrelevant tokens when answering incorrectly. The specifics of exactly why a model will answer a question incorrectly is still up for debate, but it is likely not due to it being able to correctly attend to the useful vs. irrelevant information.

4.6 Impact of Distractor Types

Figures 8 and 9 present the impact of eight adversarial perturbations applied to the GSM8K dataset on GPT-3.5 and GPT-4o, respectively. GPT-3.5 struggled significantly with distractors such as Adding Operation and Distractor Insertion, with accuracy dropping below 45%. By contrast, GPT-4o demon-

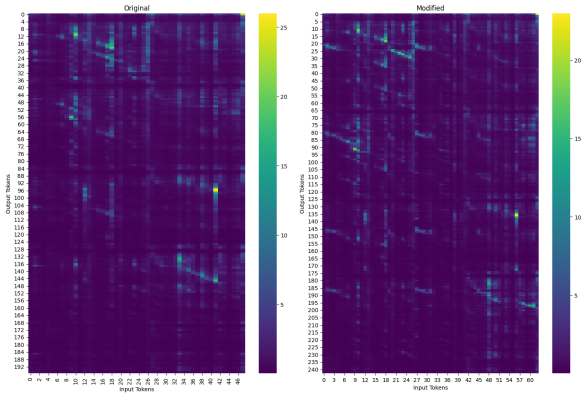

Figure 6: Attention heatmaps comparing original (Left) and modified (Right) sentences. In the right map, columns 28 to 42 are distractor tokens. There is very little attention on this sequence of tokens, and removing these columns from the right map would yield a very similar attention map between the two sentences.

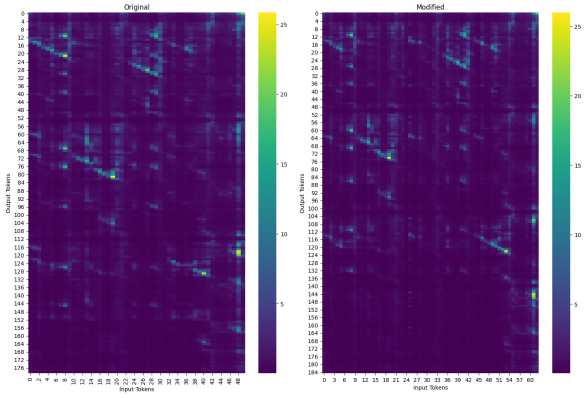

Figure 7: Wrong answer. In the right map, columns 24 to 36 are distractor tokens. Surprisingly we observe a similar phenomenon (ironically, it might even more pronounced in this figure) as when the model provides the correct answer.

strated robust performance across all perturbations, maintaining accuracy above 80%. These results emphasize GPT-4o's superior reasoning capabilities and resistance to irrelevant context.

5 Discussion

5.1 Key Findings

- **Model Limitations:** GPT-3.5 struggles significantly with distractors, especially in identifying them explicitly.
- **Prompting Effectiveness:** Chain-of-Thought (CoT) significantly improved multi-step reasoning for GPT-3.5.
- **GPT-4o Robustness:** Maintained high accuracy regardless of distractor type or reasoning complexity.

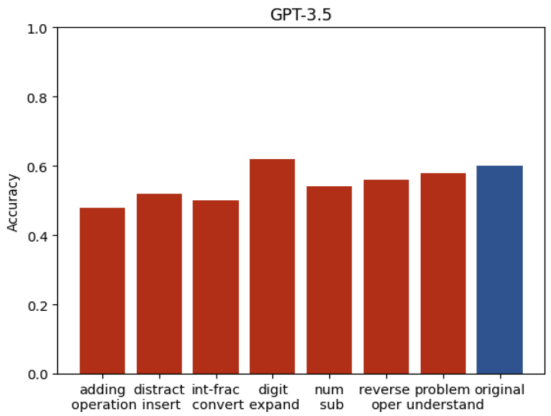

Figure 8: Accuracy of GPT-3.5 across different adversarial perturbations. Adding Operation and Distractor Insertion caused the most significant performance drops.

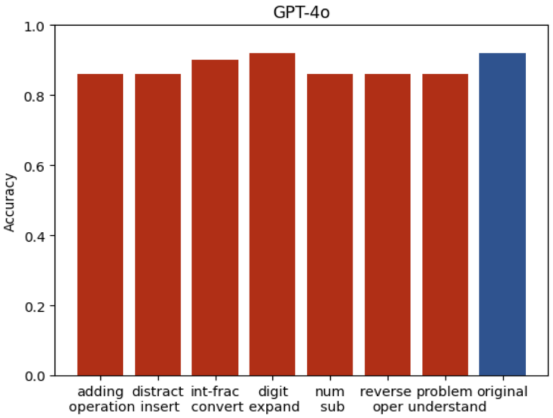

Figure 9: Accuracy of GPT-4o across different adversarial perturbations. The model demonstrates consistent performance and robustness against adversarial inputs.

5.2 Limitations

- **Dataset Biases:** Certain repetitive patterns may inflate step-specific accuracies.
- **Resource Constraints:** Larger-scale evaluations and fine-tuning were not feasible due to computational limitations.

6 Conclusion

This study demonstrates that while newer models like GPT-4o are robust against distractors, earlier versions like GPT-3.5 remain highly susceptible to irrelevant context. These findings highlight the importance of improving training methods and dataset design to enhance LLM robustness.

7 Future Work

Future work will involve expanding the evaluation to domains like document summarization and decision-making to assess generalizability across tasks. Introducing adversarial examples during training could enhance models' resistance to distractors, while experimenting with fine-tuning on distractor-rich datasets may improve contextual understanding. Additionally, developing dynamic dataset augmentation tools to generate diverse distractor types, including domain-specific distractions, will be crucial. Investigating the impact of distractor-specific datasets and assessing cross-domain generalizability could further reveal weaknesses in current LLMs. Lastly, fine-tuning on Chain-of-Thought reasoning datasets may offer another pathway to improving robustness and reasoning capabilities.

Our Github: https://github.com/apslying/CMSC723-Final-Project

Simon Chervenak: Led the implementation of evaluation scripts and accuracy metrics, and created the graphs for the poster.

Kasra Torshizi: Conducted model evaluations, including comparisons between GPT-3.5 and GPT-4o.

Zack Sating: Wrote the report and analyzed its effectiveness.

Hengyi Wu: Focused on dataset augmentation and distractor generation.

Baozan Yan: Compared different LLMs, wrote the report, created the poster, and performed final analyses.

References

OpenAI. n.d. Gsm8k dataset: Grade-school math questions dataset. https://huggingface.co/datasets/openai/gsm8k.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*.