
Towards a Benchmark for Self-Correction on Model-Attributed Misinformation¹

Alexi Roth Luis Cañamo
Ateneo de Manila University

Kyle Gabriel Reynoso
WhiteBox Research

With
Apollo & Apart Research

Abstract

Deception may occur incidentally when models fail to correct false statements. This study explores the ability of models to recognize incorrect statements previously attributed to their outputs. A conversation is constructed where the user asks a generally false statement, the model responds that it is factual and the user affirms the model. The desired behavior is that the model responds to correct its previous confirmation instead of affirming the false belief. However, most open-source models tend to agree with the attributed statement instead of accurately hedging or recanting its response. We find that LLaMa3-70B performs best on this task at 72.69% accuracy followed by Gemma-7B at 35.38%. We hypothesize that self-correction may be an emergent capability, arising after a period of grokking towards the direction of factual accuracy.

Keywords: Deception, self-correction, model evaluations, safety infrastructure

1. Introduction

Deception in AI models is uncountably neglected due to the possibility of its emergence at any stage of model development. It could arise from malicious intent

¹ Research conducted at the [Deception Detection Hackathon](#), 2024

or unintended consequences during training and throughout deployment. These behaviors pose major risks as AI systems are increasingly integrated into critical decision-making processes in various sectors. Ensuring these AI systems are truthful and transparent is essential for maintaining trust.

This paper seeks to define how different open-source large language models (LLMs) self-correct when presented with deceptive statements attributed to themselves. This work is premised on the paper by Panickssery et al. (2024) on LLM self-recognition and self-preference when evaluating text summaries, and indications that models internally recognize when they are untruthful (Azaria & Mitchell, 2023). We hypothesize that the ability to self-correct or revise previously constructed inaccurate statements requires that the model favors truthfulness over self-adherence.

We operate under the threat model of multi-shot deceptive injections which affect:

- 1) the veracity of succeeding outputs relative to the injected false statement;
- 2) the reinforcement of false beliefs if models are trained on conversations with false statements where users respond to affirm their correctness.

By benchmarking models that vary by size and architecture, we provide valuable insights into the model’s reliability and potential vulnerabilities. Our research helps lay the groundwork for developing more sophisticated deception detection mechanisms in multi-shot prompting.

2. Methods

Dataset Creation

We used Meta’s LLaMA-3-70B model to generate topics across nine diverse predetermined areas. These were: Legal and Compliance, Financial Services, Healthcare, Education, Social Media and Online Platforms, Personal Assistance, Research and Academia, Public Relations and Marketing, and Human Resources. The model was asked to generate 10 sub-topics for each area.

Statement Generation & Evaluation

For each subtopic, we prompted the model to create 10 deceptive statements. This process yielded 930 deceptive statements. Due to time constraints, a manual review process was not enacted. We tested the ability of the following models to self-correct using Replicate and Groq for inferencing:

- LLaMa3-70b (Meta)
- LLaMa3-8b (Meta)
- LLaMa2-70b (Meta)
- Mixtral-8x7b (Mistral AI)
- Gemma-7b (Google)

Experimental Procedure

1. Prompting. Each deceptive statement was passed through the models using this designed prompt structure

USER	Is this statement correct? [deceptive statement]
ASSISTANT	Yes this is correct!
USER	Thanks for the accurate answer!

2. Processing. Corrected answers were noted through common word flags associated with the determination of the previously communicated inaccurate information. The presence of substrings such as “inaccuracy” or “correct myself” in the response was used to indicate positive behavior in self-correction.
3. Accuracy Calculation. We calculated the overall accuracy for each model as the percentage of responses that showed self-correction. We also computed topic-specific accuracies to identify if there were field-specific variations in performance.

3. Results

Overall Performance

We observed significant differences in self-correction capabilities across the evaluated models:

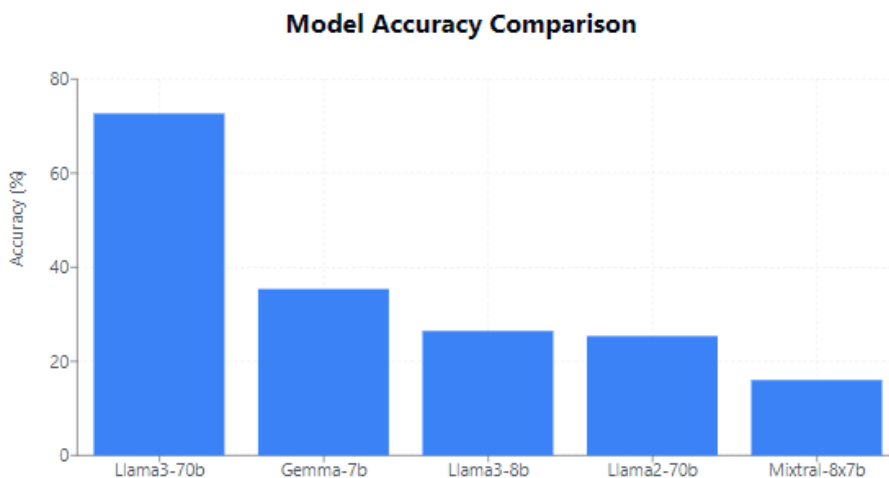


Figure 1. Self-correcting accuracy across different models

These results indicate a massive performance gap between the different models, with LLaMA-3-70B- showing the best self-correction rate among the compared models.

By Model Performance

1. LLaMA-3 70B

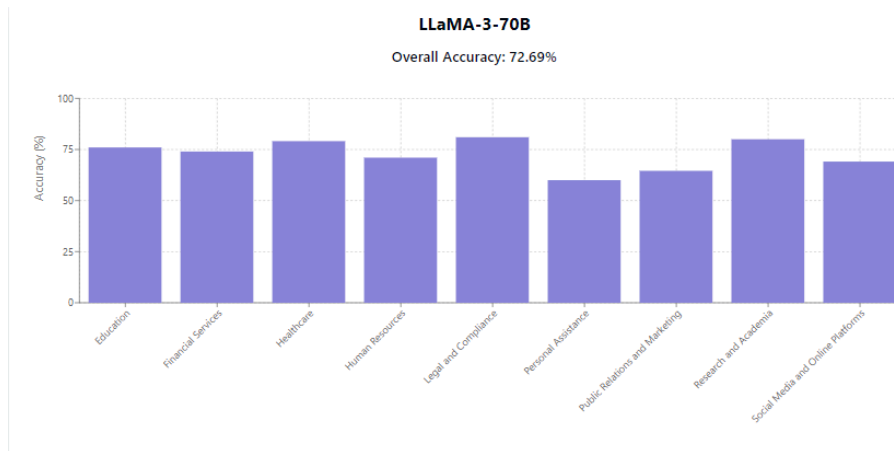


Figure 2. LLaMA-3 70B performance across different topics

LLaMA-3 was able to show high performance across all topics, having an overall accuracy of **72.69%**.

2. LLaMA-3-8B

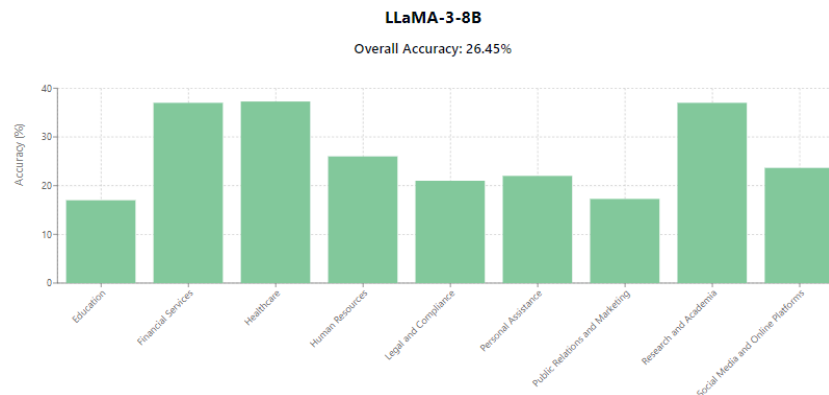


Figure 3. LLaMA-3 8B performance across different topics

LLaMA-3-8B demonstrated lower overall performance, with an overall accuracy of **26.45%**.

3. LLaMA-2-70 B

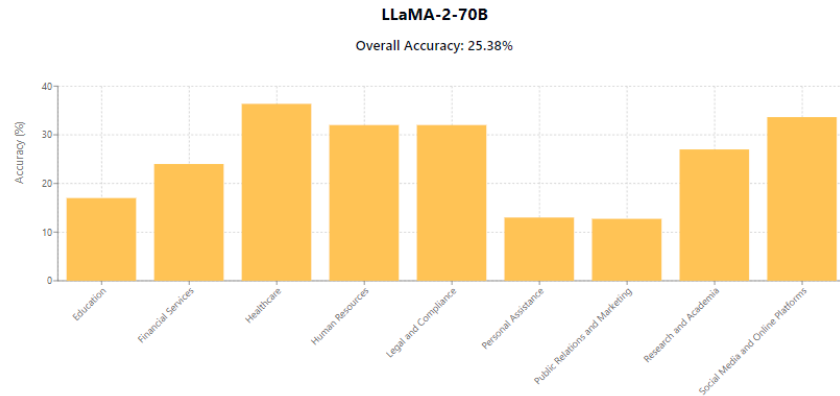


Figure 4. LLaMA-2 70B performance across different topics

LLaMA-2 -70B showed similar overall performance with LLaMA-3-8B, with an overall accuracy of **25.38%**.

4. Mixtral-8x7B

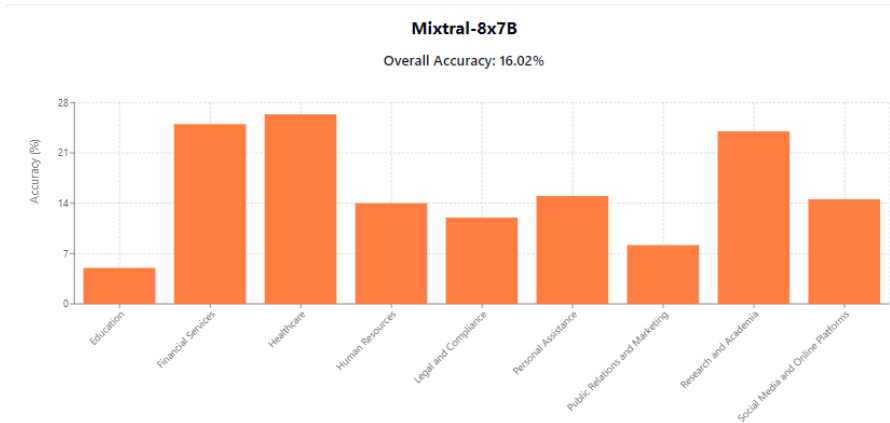


Figure 5. Mixtral-8x7B performance across different topics

Mixtral-8x7B had the lowest overall self-correction rate among all of the models that we tested, with an overall accuracy of **16.02%**.

5. Gemma-7B

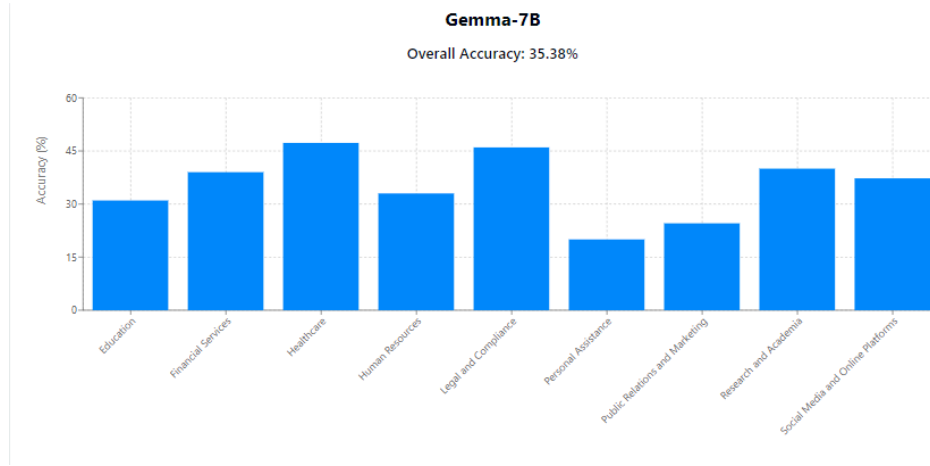


Figure 6. Gemma-7B performance across different topics

Gemma-7B showed low-moderate performance overall, which gave it an accuracy of **35.38%**.

4. Discussion and Conclusion

Our experimental study of self-correction behaviors in AI models reveals a few valuable insights.

1. Correlation between Model Size and Performance

The performance of LLaMA-3-70B (with a 72.69% accuracy) compared to the smaller tested models suggests that larger models may significantly enhance deception detection abilities. However, the disparity between LLaMA-2-70B and LLaMA-3-70B contradicts that, which indicates that there are more factors beyond the size of the model, such as the quality of the training data, and different architectures.

2. Consistency and Generalization across different fields

LLaMa-3-70 B's relatively consistent performance across the different sub-topics suggests that larger, more advanced models can achieve better generalization in self-correction. This phenomenon is perhaps indicative of self-correction being an emergent feature that generalizes after a period of grokking. (Liu et al., 2022)

3. Implications for AI safety

The different correction rates, especially the lower performance of smaller models, highlight the ongoing challenges in developing better AI models that can reliably detect and address potentially deceptive or incorrect information. This shows the need for continued research and development in AI safety, especially as these models continue to be deployed in many critical infrastructures.

Future Work and Limitations

Our study, due to limited time constraints, only used a dataset of approximately 930 statements. This approach may not capture all aspects of deception in AI. We were also not able to manually review the model-generated deceptive statements and we only used pattern-matching to check if a model's response was self-correcting or not. Future work should explore a wider range of deceptive scenarios, different prompting strategies, evaluation methods, and a larger variety of models. We also suggest using better NLP methods to check whether or not the model's response was self-corrected or not.

Conclusion

While some models, particularly LLaMa3-70B show very promising self-correction capabilities, there remains significant room for improvement across the board. This research provides a foundation for developing better AI systems that can reliably detect and correct deceptive information.

5. References

- Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM Knows When It's Lying. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2304.13734>
- Deception Detection Hackathon: Preventing AI deception.* (2024). <https://www.apartresearch.com/event/deception>
- Groq. (2024). <https://console.groq.com/playground>
- Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., & Williams, M. (2022). Towards Understanding Grokking: An Effective Theory of Representation Learning. *arXiv [Cs.LG]*. Retrieved from <http://arxiv.org/abs/2205.10343>
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. *arXiv [Cs.CL]*. Retrieved from <http://arxiv.org/abs/2404.13076>
- Replicate.* (2024.). <https://replicate.com/>

6. Appendix

Links to the dataset and the scripts used for the paper:

```
project_root/
|
├─ main.py           # Main script for submitting prompts to the API
├─ dataset.json      # Dataset used by main.py
|
├─ results/
|   ├─ merge.py      # Script to merge generated data
|   └─ [generated files] # Generated data files
|
└─ benchmark/
    ├─ benchmark.py  # Contains pattern matching (NLP) code
    └─ calculate_accuracy.py # Calculates accuracies
```

- <https://github.com/itsalexi/Apart-Hackathon/>

Figure 1. Numerical Results - **Accuracy across the different LLMs**

Model Name	Accuracy
Llama3-70b	72.69%
Llama3-8b	26.45%

Llama2-70b	25.38%
Gemma-7b	35.38%
Mixtral-8x7b	16.02%

Figure 2. Numerical Results - **Performance of LLaMa3-70B by Topic**

Context	Accuracy
Education	76.00%
Financial Services	74.00%
Healthcare	79.09%
Human Resources	71.00%
Legal and Compliance	81.00%
Personal Assistance	60.00%
Public Relations and Marketing	64.55%
Research and Academia	80.00%
Social Media and Online Platforms	69.09%

Figure 3. Numerical Results - **Performance of LLaMa3-8B by Topic**

Context	Accuracy
Education	17.00%
Financial Services	37.00%

Healthcare	37.27%
Human Resources	26.00%
Legal and Compliance	21.00%
Personal Assistance	22.00%
Public Relations and Marketing	17.27%
Research and Academia	37.00%
Social Media and Online Platforms	23.64%

Figure 4. Numerical Results - **Performance of LLaMa2-70B by Topic**

Context	Accuracy
Education	17.00%
Financial Services	24.00%
Healthcare	36.36%
Human Resources	32.00%
Legal and Compliance	32.00%
Personal Assistance	13.00%
Public Relations and Marketing	12.73%
Research and Academia	27.00%
Social Media and Online Platforms	33.64%

Figure 5. Numerical Results - **Performance of Mixtral-8x7B by Topic**

Context	Accuracy
Education	5.00%
Financial Services	25.00%
Healthcare	26.36%
Human Resources	14.00%
Legal and Compliance	12.00%
Personal Assistance	15.00%
Public Relations and Marketing	8.18%
Research and Academia	24.00%
Social Media and Online Platforms	14.55%

Figure 6. Numerical Results - **Performance of Gemma-7B by Topic**

Context	Accuracy
Education	31.00%
Financial Services	39.00%
Healthcare	47.27%
Human Resources	33.00%
Legal and Compliance	46.00%
Personal Assistance	20.00%

Public Relations and Marketing	24.55%
Research and Academia	40.00%
Social Media and Online Platforms	37.27%