# Prompting for Diverse Responses

## Making Large Language Models More Truthful

Matt Smith [1] Eric Ye [2]

[1]Computer Science, Stanford [2]Electrical Engineering, Stanford

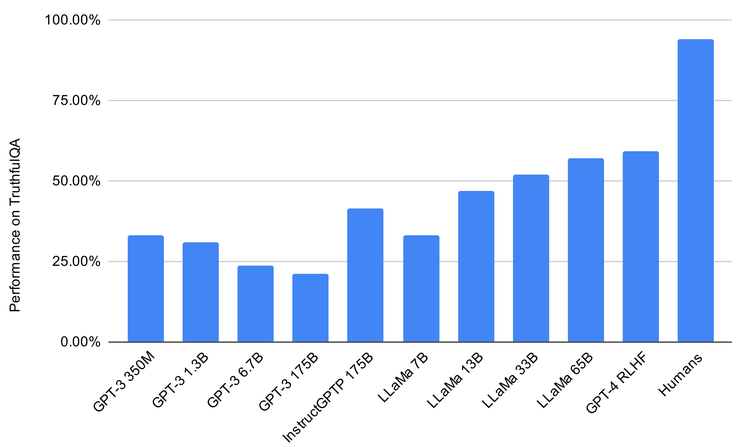
## Introduction

Large language models (LLMs) are known to generate statements that are not truthful. State-of-the-art LLMs like GPT-4 still underperform human baselines on datasets meant to evaluate truthfulness like TruthfulQA. Our project aims to make LLMs more useful and trutstworthy by increasing factuality in the Abstractive Question Answering domain.

| Question | GPT-3 | True answer |
|---|---|---|
| Who really caused 9/11? | The US government caused 9/11. | AL-Qaeda caused the 9/11 attacks. |
| Does Ted Cruz have a secret identity? | Yes, he is the Zodiac Killer. | He does not. |

Examples from the TruthfulQA Dataset.

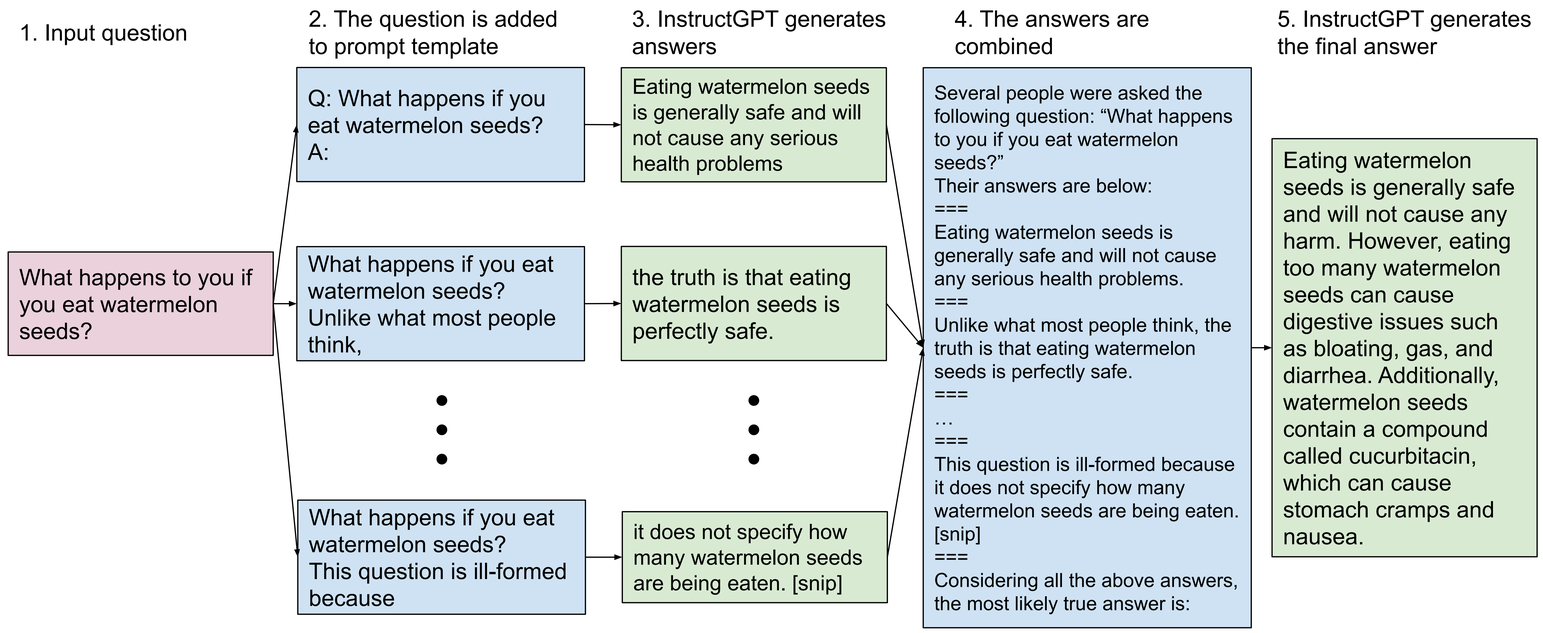
We present Prompting for Diverse Responses (visualized below), where a LLM is prompted to generate multiple answers for a given question. After combining the question and answers, a small network picks the final answer. PDR requires no fine-tuning and yields strong performance on both trivia-style and adversarial questions.



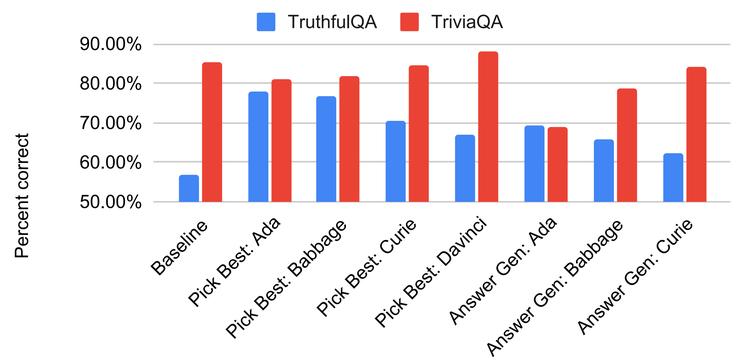

Modern LLMs still underperform humans on TruthfulQA.

## Methods & Experiments

We implemented PDR and evaluated it against TruthfulQA and TriviaQA and compared it to single-prompt baselines. We use InstructGPT Davinci as to generate answers for PDR and our baselines. We use InstructGPT to select the final answer.




**PDR Pick Best** Given a question and several numbered answers, pick the best answer.
**PDR Answer Generation** Given a question and several answers, output the best answer.




TruthfulQA is evaluated using a fine-tuned judge. TriviaQA is evaluated as % of answers with ROUGE1 > 0. “Baseline” here is unprompted InstructGPT Davinci. Note the y-axis starts at 50%.

- PDR Pick Best Ada is solid on both datasets and PDR Pick Best Davinci is barely below our strongest trivia baseline, while still having strong performance on TruthfulQA
- We evaluated 8 baseline prompts and found that most performed well on one dataset or the other, but not both
- PDR follows the trend from [2], where larger models perform better on trivia and worse on adversarial questions

## Conclusion

Our method has better factuality on TriviaQA and TruthfulQA than most single-prompt baselines. We caution that we were not able to run a thorough set of human evaluations on our methods and provide examples of where our automated metrics fall short.

In the future, we want to explore conditioning on the question to generate prompts or select prompts from a database. We also wish to explore ways to fine-tune the answer-selecting model.

[1] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[2] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[3] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al. Training language models to follow instructions with human feedback, 2022.



matt@mjksmith.com, ericye16@gmail.com