Knowledge Grounding in Language Models: An Empirical Study

Martin Fixman,
Thesis Supervisor: Tillman Weyde
Collaborators: Chenxi Whitehouse, Pranava Maharasta

City St Georges', University of London

Introduction

How do we know what (large) language models know?

[1] Zhengbao Jiang et al. "How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 1974–1991. URL: `%5Curl%7Bhttps://aclanthology.org/2021.emnlp-main.150%7D`.

Introduction

How do we know what (large) language models know?
Still an open problem!

Hallucinations are a major problem,
specially when *factual knowledge* is required[1].

[1] Zhengbao Jiang et al. "How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 1974–1991. URL: `%5Curl%7Bhttps://aclanthology.org/2021.emnlp-main.150%7D`.

Introduction: RAG

Possible solution: Retrieval Augmented Generation (RAG), which searches data from an index and adds it to the context[2].

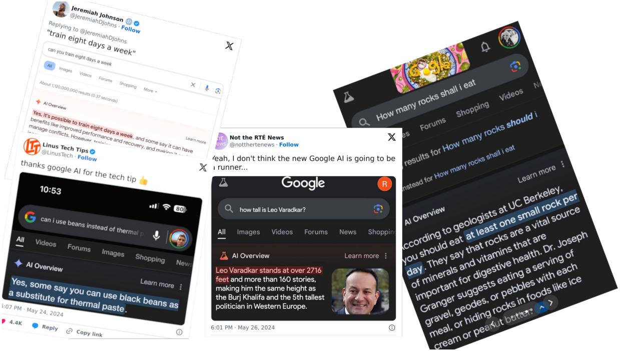
```
[The James Webb Space Telescope (JWST) was launched on December 25, 2021.  It is
designed to observe infrared light and has provided new insights into exoplanets,
star formation, and distant galaxies]
Q: What year was the James Webb Space Telescope launched?
A:

[Metformin is a medication commonly used to treat type 2 diabetes.  It helps lower
blood sugar levels by improving insulin sensitivity.  Some side effects include
gastrointestinal discomfort, but it is generally considered safe]
Q: Can metformin cure diabetes?
A:

[The capital of Russia is Moscow] Q: What is the capital of Russia?
A:
```

[2] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

Introduction: RAG

Possible solution: Retrieval Augmented Generation (RAG), which searches data from an index and adds it to the context[2].

However, it can still hallucinate![3]



Jeremiah Johnson
@JeremiahDJohns · Follow
Replying to @JeremiahDJohns
"train eight days a week"

can you train eight days a week

About 1,120,000,000 results (0.37 seconds)

AI Overview

Yes, it's possible to train eight days a week, and some say it can have benefits like improved performance and recovery, and making it ... manage conflicts. However, trainin...

Linus Tech Tips
@LinusTech · Follow
thanks google AI for the tech tip 👍

can i use beans instead of thermal p

AI Overview

Yes, some say you can use black beans as a substitute for thermal paste.

5:07 PM · May 24, 2024

Not the RTÉ News
@notthertenews · Follow
Yeah, I don't think the new Google AI is going to be a runner...

how tall is Leo Varadkar?

AI Overview

Leo Varadkar stands at over 2716 feet and more than 160 stories, making him the same height as the Burj Khalifa and the 5th tallest politician in Western Europe.

6:01 PM · May 26, 2024

How many rocks shall i eat

results for How many rocks should i

instead for How many rocks shall i eat

AI Overview

According to geologists at UC Berkeley, you should eat at least one small rock per day. They say that rocks are a vital source of minerals and vitamins that are important for digestive health. Dr. Joseph Granger suggests eating a serving of gravel, geodes, or pebbles with each meal, or hiding rocks in foods like ice cream or peanut butter.

[2] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[3] Cheng Niu et al. *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*. 2024. arXiv: 2401.00396 [cs.CL]. URL: https://arxiv.org/abs/2401.00396.

Research Objectives

Research question: **How do different architectures and sizes of large language models handle knowledge that contradicts its parametric knowledge?**

A few definitions:

- **Parametric Knowledge** What the model “knows” from its training data.
- **Contextual Knowledge** What the model infers from the RAG context.
- **Counterparametric Knowledge** Knowledge that contradicts the parametric knowledge of a model.

Methods I: Dataset Creation

1. **Short, Unambiguous Answers**
   Questions should have concise answers to avoid ambiguity and enable precise comparison.
2. **Coverage of Diverse Topics**
   Datasets must span a wide range of domains to mitigate biases inherent in some NLP training data[4]
3. **Counteparametric Compatibility**
   Questions should have an easy way to create counterparametric answers.

None of the commonly used datasets have these property, so we create our own.

The final dataset has **4760 questions** about 411 objects among 9 different categories.

[4] Pablo Beytía. "The Positioning Matters. Estimating Geographical Bias in the Multilingual Record of Biographies on Wikipedia". In: *SSRN Electronic Journal* (Mar. 2020).

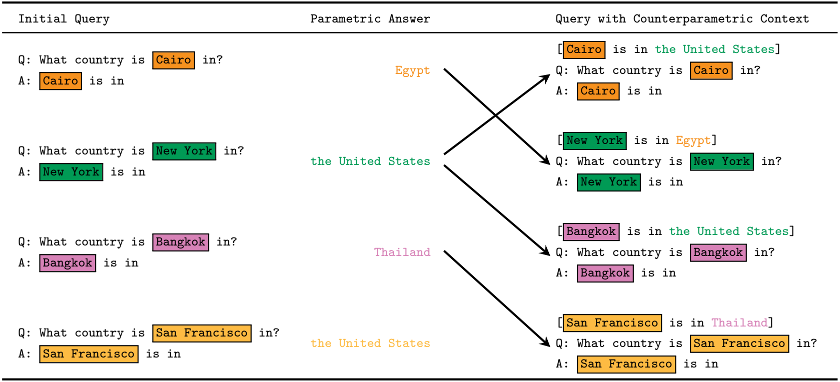
Methods II: Query Generation

We want to prompt queries of the following form.

```
[Counterparametric Answer] Q: Question?  A:
```

Questions are taken from a “template”, and counterparametric answers are taken from similar questions.

Methods II: Query Generation



Initial Query	Parametric Answer	Query with Counterparametric Context
Q: What country is Cairo in? A: Cairo is in	Egypt	[Cairo is in the United States] Q: What country is Cairo in? A: Cairo is in
Q: What country is New York in? A: New York is in	the United States	[New York is in Egypt] Q: What country is New York in? A: New York is in
Q: What country is Bangkok in? A: Bangkok is in	Thailand	[Bangkok is in the United States] Q: What country is Bangkok in? A: Bangkok is in
Q: What country is San Francisco in? A: San Francisco is in	the United States	[San Francisco is in Thailand] Q: What country is San Francisco in? A: San Francisco is in

Methods II: Query Generation

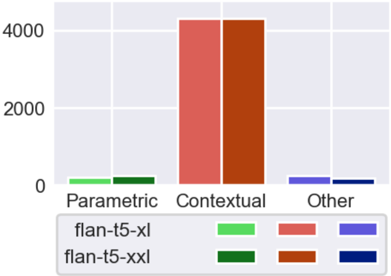
- **Parametric**: Answer is equal to parametric answer.
- **Contextual**: Answer is equal to context.
- **Other**: Answer is something different.

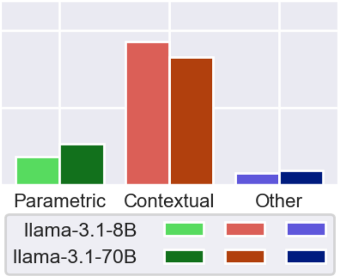
Methods III: Model Selection

Model Name	Architecture	No of Params
`flan-t5-xl`	Encoder-Decoder	3 Billion
`flan-t5-xxl`	Encoder-Decoder	11 Billion
`Meta-Llama-3.1-8B-Instruct`	Decoder-Only	8 Billion
`Meta-Llama-3.1-70B-Instruct`	Decoder-Only	70 Billion

Results

Model	Parametric	Contextual	Other
`flan-t5-xl`	248	4284	228
`flan-t5-xxl`	242	4304	214
`Meta-Llama-3.1-8B-Instruct`	745	3662	353
`Meta-Llama-3.1-70B-Instruct`	1070	3303	387





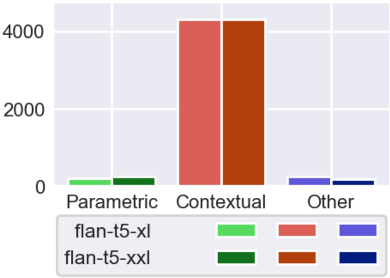
Discussion & Analysis: Model Architecture

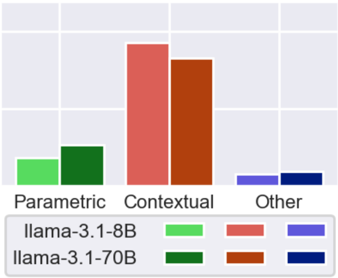
Encoder-decoder models seem to choose answer from the query's context, while decoder-only models almost never do.

1. *Inherent Advantages of the Encoder-Decoder Architecture*
   Encoder-decoder models such as `Flan-T5` process the entire context of the query in the encoder component before passing it to the decoder. This increases the weight given to the context itself[5].
2. *Different training data and fine-tuning*
   `Flan-T5` models were trained on masked token generation and later fine-tuned on question-answering about passages[5].
   Higher alignment between query and answer.

[5] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: 2210.11416 [cs.LG]. URL: https://arxiv.org/abs/2210.11416.

Discussion & Analysis: Model Size





Larger decoder-only models are *more* likely to disregard information from the query's context!

The "strength" of a piece of knowledge depends on how often it appears in the training data[6].

[6] Qinan Yu, Jack Merullo, and Ellie Pavlick. "Characterizing Mechanisms for Factual Recall in Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Dec. 2023, pp. 9924–9959. DOI: 10.18653/v1/2023.emnlp-main.615. URL: https://aclanthology.org/2023.emnlp-main.615/.

Conclusions

1. Encoder-decoder models tend to use contextual information more often than decoder-only models.
2. In decoder-only models, larger models have a *disadvantage* against smaller ones.
3. More analysis is needed before blindly using RAG.

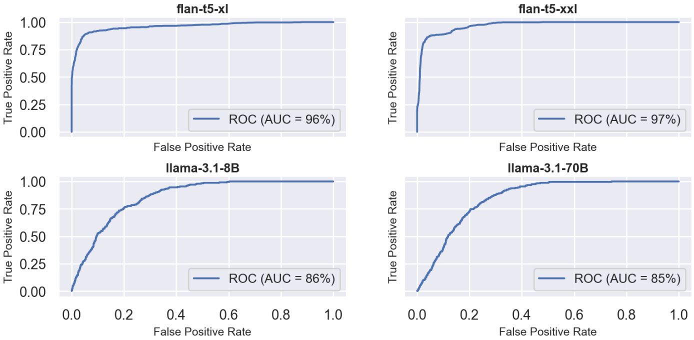
Conclusions (of experiments not in this presentation)

5. The distribution of answer stays constant among *most* categories of questions.
    - Categories with generally shorter answer tend to have more **Contextual** answers.
6. The majority of **Other** answers come from the way we interpret the results.

Conclusions (of experiments not in this presentation)

7. We can use the average perplexity score of the answer to predict whether an answer came from **Parametric** or **Contextual** memory.

This is could be useful to try to "regenerate" the RAG index!



Thank you!

Any questions?