Introduction.
Large Language Models (LLMs) offer state-of-the-art performance on many tasks. However, hallucinations remain problematic in mission-critical contexts.

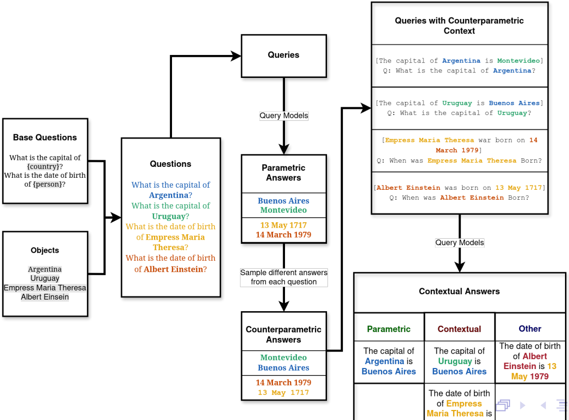
Motivation and Research Question

Motivation.
Retrieval-Augmented Generation (RAG) is a promising way to mitigate hallucinations by providing external context to an LLM.
Research Question.
How do LLMs respond if the context provided contradicts what they have memorized in their parameters?

Method: Framework Overview

Here is an overview of the experimental setup.

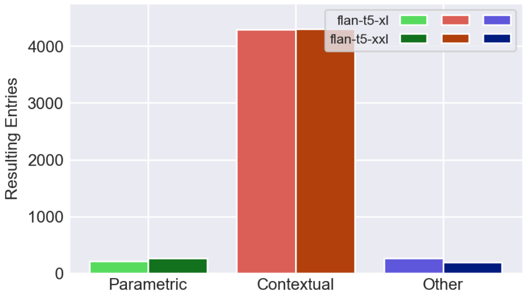
- Generate a diverse dataset of short-answer questions.
- Query the model without extra context to get a **Parametric** answer.
- Add a *counterparametric* context that contradicts the original model answer.
- Re-query with the new contradictory context.
- Compare the new response against parametric vs. contextual data.



Results: Parametric vs. Contextual

We tested four models: two Seq2Seq (Flan-T5) and two Decoder-only (Llama).

- **Contextual** answers dominate in Seq2Seq models.
- Decoder-only models more often ignore contradictory context and revert to **Parametric** knowledge.



Discussion: Model Architecture and Size

Seq2Seq models (encoder-decoder) appear more sensitive to external context.

- **Flan-T5-XL and Flan-T5-XXL**: minimal difference in using context despite large size gap.
- **Llama-8B vs. Llama-70B**: bigger model reverts to parametric memory more often.

Conclusion.
Bigger Decoder-only models are more likely to trust memorized facts over a contradictory context, while Seq2Seq architectures generally rely on provided context.

Future Work

Refine string-comparison to handle partial rephrasings.
Extend experiments to:

- RAG-specific models like Atlas or Retro.
- Fine-tuning large language models to better trust contradictory context.
- Using perplexity signals to detect hallucinations and selectively re-query the retriever.

References

Thank you. Questions?