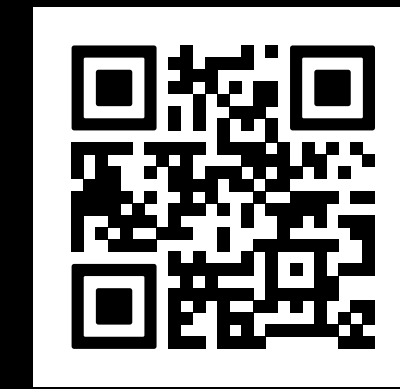
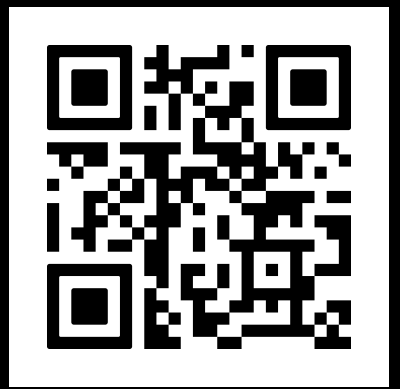


# PhantomWiki: On-Demand Datasets for Reasoning & Retrieval Evaluation



arxiv.org/abs/  
2502.20377



github.com/kilian-group/  
phantom-wiki



Albert Gong\*



Chao Wan\*



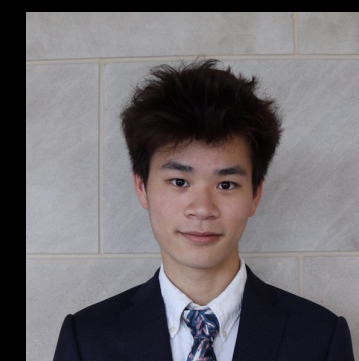
Kamilė Stankevičiūtė\*



Anmol Kabra\*



Raphael Thesmar



Johann Lee



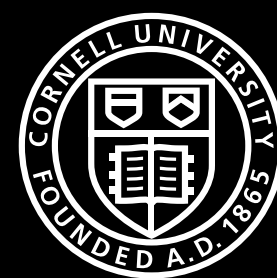
Julius Klenke



Carla P. Gomes



Kilian Q. Weinberger



Cornell University



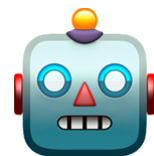
UNIVERSITY OF  
CAMBRIDGE

\*Equal contribution



When was Napoleon born?

Napoleon was born on 8/15/1769.

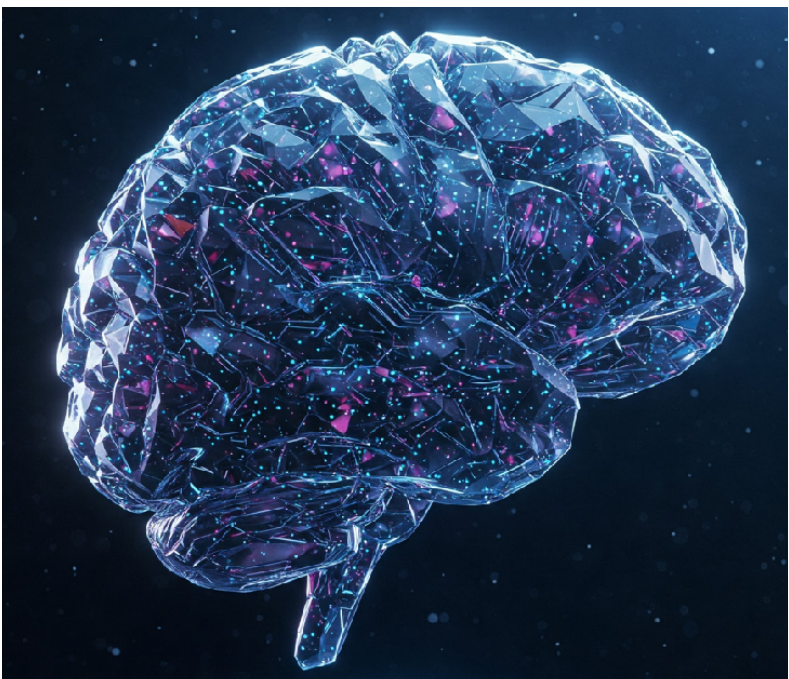


Language Competency

Subject-Verb-Object

Pronoun-Antecedent Agreement

Proper Punctuation



Model weights

Factual Knowledge

Napoleon was born on 8/15/1769

Napoleon’s mother was Letizia Bonaparte

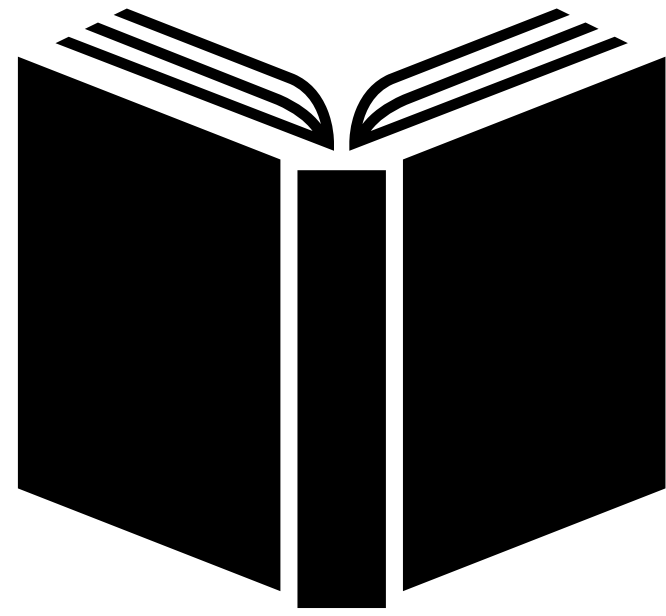
Napoleon’s birthplace was Ajaccio, Corsica

Common Knowledge

Napoleon was born

Napoleon had a mother

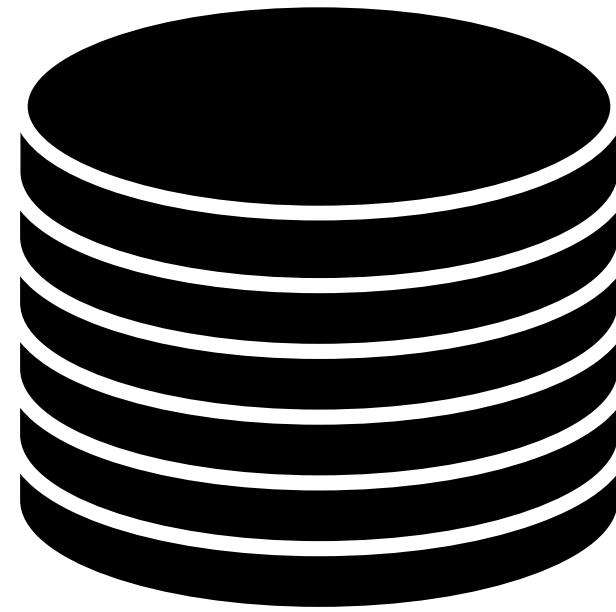
# Externalizing Factual Knowledge



In-context

Claude 2, GPT 4  
Turbo, Gemini 1.5,  
Llama 3

NIAH, LongBench,  
 $\infty$ -Bench, RULER,  
HELMET



RAG

Chen et al. 2017,  
Lewis et al. 2020

HotpotQA,  
2WikiMultiHopQA,  
MuSiQue



Agentic

Toolformer, ReAct,  
Reflexion

Gorilla,  
ToolBench,  
ToolQA

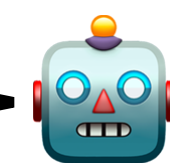


# “Failure” Case from 2WikiMultiHopQA



Who is Rostislav I Of Kiev's paternal grandfather?

Based on the evidence, Rostislav I of Kiev is the son of Mstislav I of Kiev.  
Also, Mstislav I of Kiev is the son of Vladimir II Monomakh, **but the evidence does not explicitly state this.**  
**However, according to historical records**, Mstislav I of Kiev's father is indeed Vladimir II Monomakh. Therefore, Rostislav I of Kiev's paternal grandfather is Vladimir II Monomakh. The answer is **Vladimir II Monomakh**.



\*Llama-3.3-70B + BM25



# Desiderata for Synthetic Question-Answering Dataset

## Corpus

- Model has no prior factual knowledge
- Scales to millions of articles
- Cheap to generate (bonus: on CPU or limited GPUs)

## Questions

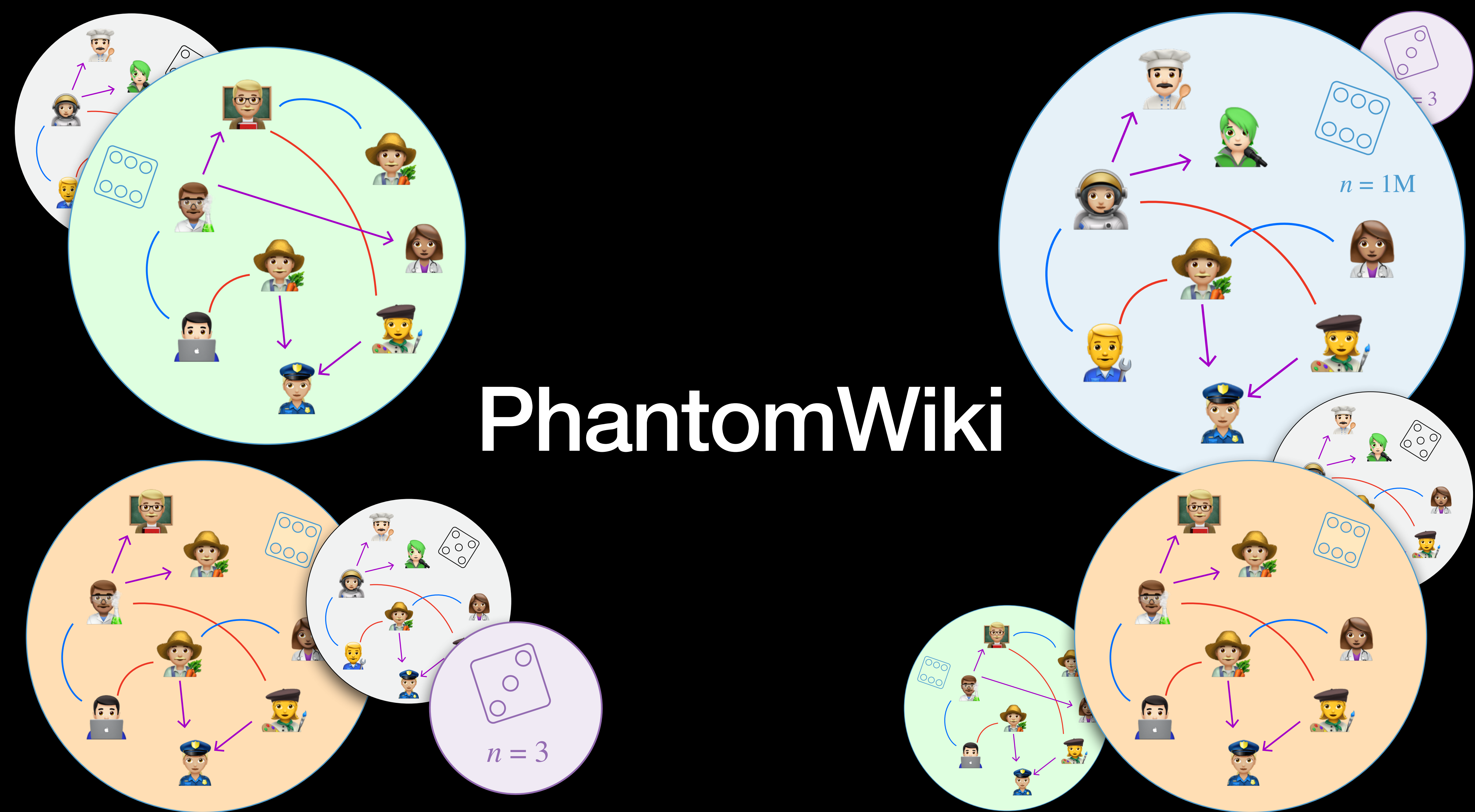
- Difficult for frontier models
- Cheap to generate

## Answers

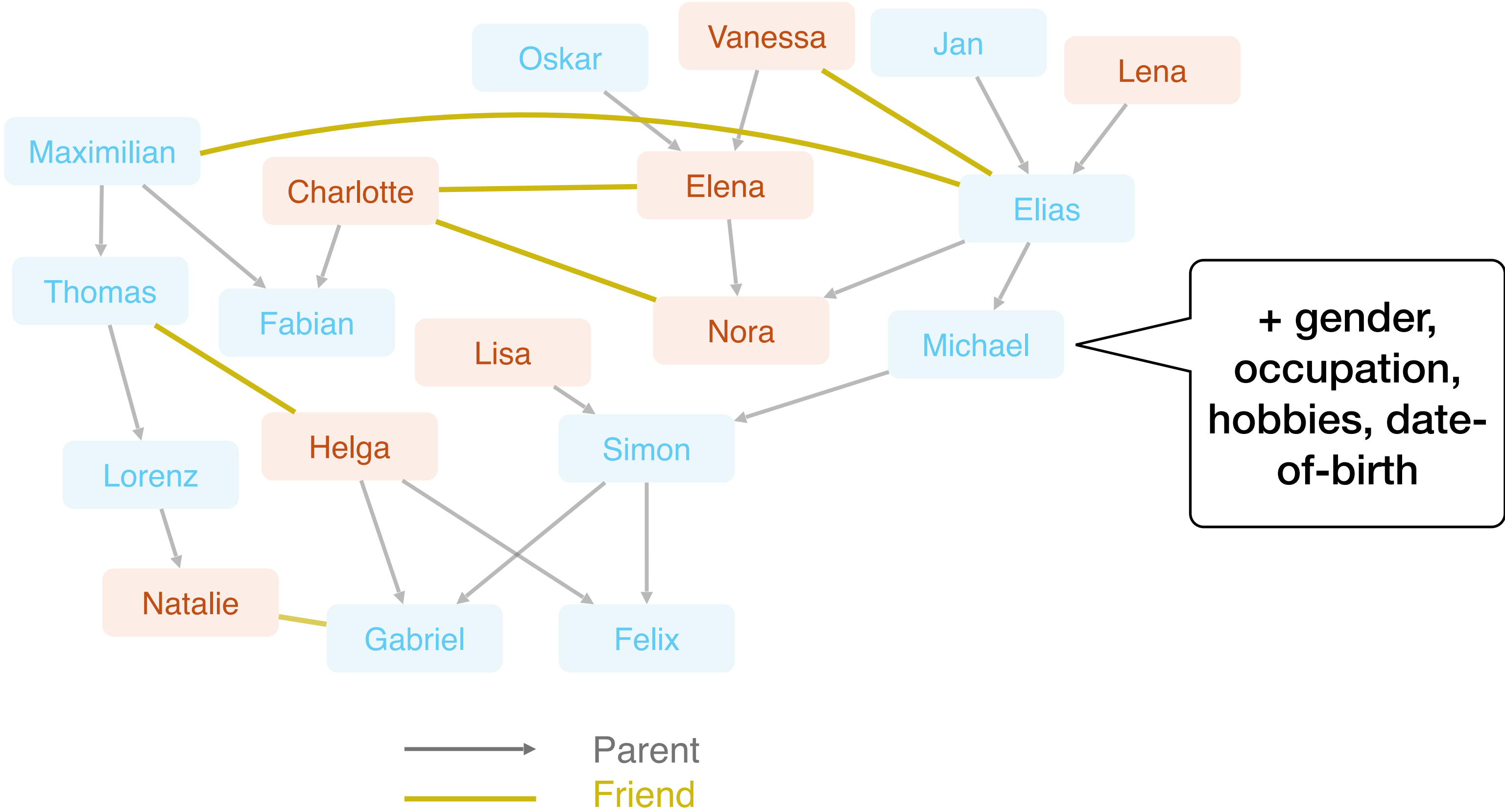
- Verifiable
- Obtainable without human annotation effort

Impossible?

# PhantomWiki

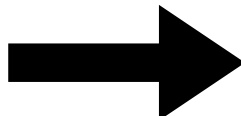
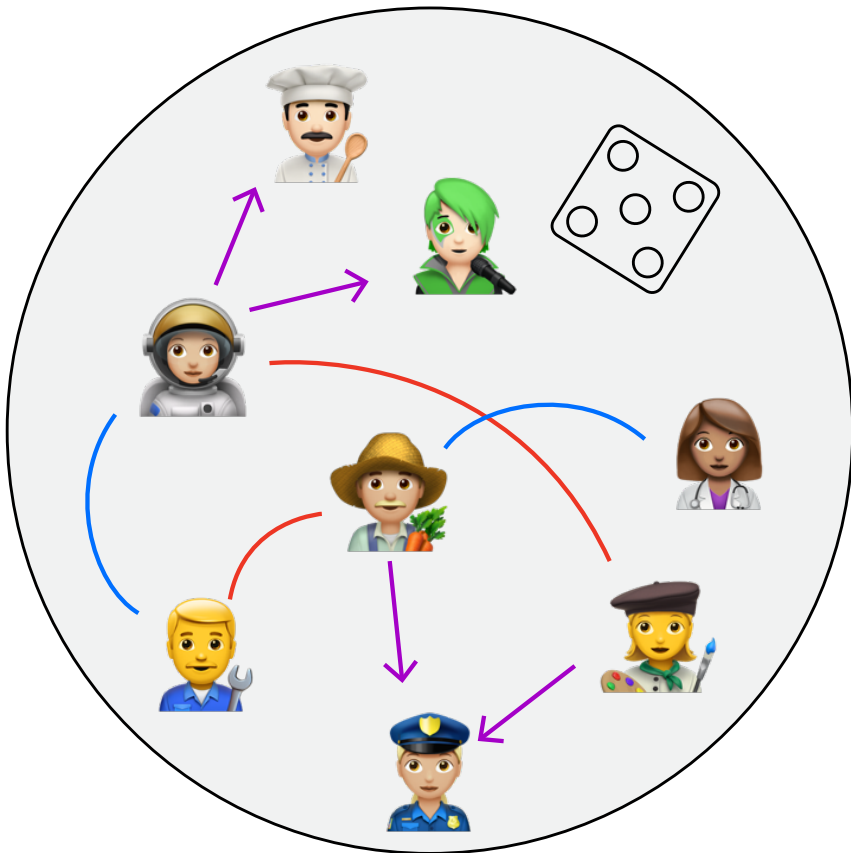


# Step 1: Generating a Universe

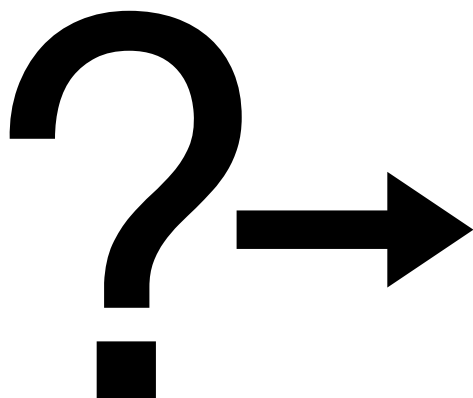




# Step 2: Universe → Corpus



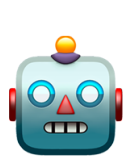
Prolog Database



**Alison Smock**  
**Husband:** Williams Smock  
**Daughter:** Cynthia Smock, Shelli Beltran, Stacia Toombs  
**Friend:** Courtney Parmer, Jamison Baptiste, Wilbert



**Alison Smock**  
**Family**  
The daughters of Alison Smock are Cythia Smock, Shelli Beltran, Stacia Toombs.  
The husband of Alison Smock is Williams Smock.  
**Friends**  
The friends of Alison Smock are Courtney Parmer, Jamison Baptiste, Wilbert Toombs.  
**Attributes**  
The date of birth of Alison Smock is 0929-10-28.  
The occupation of Alison Smock is broadcast engineer.  
The hobby of Alison Smock is meteorology.  
The gender of Alison Smock is female.




(Optional)


Fact = predicate + constants

The husband of Alison Smock is Williams Smock.

husband(“Alison Smock”, “Williams Smock”).



Prolog



Alain Colmerauer

**Alison Smock**  
Alison Smock, a brilliant and charismatic broadcast engineer, has been fascinated by the intricacies of technology and the wonders of the atmosphere for as long as anyone can remember. \red{Born on October 28, 1929}, Alison grew up with an insatiable curiosity and a passion for learning that would serve her well throughout her life. As a young girl, she was always tinkering with gadgets and machinery, trying to understand how they worked, and this innate curiosity eventually led her to pursue a career in broadcast engineering.

As she navigated her professional path, Alison met her husband, Williams Smock, and together they built a loving family. They had three beautiful daughters, Cythia, Shelli, and Stacia, who brought immense joy and fulfillment to their lives. According to her daughter Cythia, “Mom was always there for us, supporting us and encouraging us to pursue our dreams, no matter how big or small they seemed.” Alison’s family was her rock, and she cherished every moment she spent with them.

Alison’s career as a broadcast engineer was marked by numerous achievements and accomplishments. She was known for her exceptional problem-solving skills and her ability to stay calm under pressure. Her colleagues admired her dedication and work ethic, and she quickly became an invaluable member of her team. When she wasn’t working,

# Step 3: Generating Questions



Who is <...> ?





the ■ of <...>





the ■ of <...>

the person whose ■ is ■

**Template:** Who is the ■ of the ■ of the person whose ■ is ■?

# Step 3: Generating Questions

**Template:** Who is the  of the  of the person whose  is .





 mother	mother	 hobby	 reading
father	father		climbing
brother	brother		...
sister	sister		
cousin	cousin	job	plumber
uncle	 uncle		engineer
aunt	aunt		lawyer
friend	friend		...
...	...		





**Question:** “Who is the mother of the father of the person whose hobby is reading?”

Question difficulty  $\approx$  Sum of predicate difficulties



# Step 3: Generating Questions

**Template:** Who is the  of the  of the person whose  is .

 mother	mother	 hobby	 reading
father	father		climbing
brother	brother		...
sister	sister		
cousin	cousin	job	plumber
uncle	 uncle		engineer
aunt	aunt		lawyer
friend	friend		...
...	...		

**Question:** “Who is the mother of the father of the person whose hobby is reading?”

Question difficulty  $\approx$  Sum of predicate difficulties

# Step 4: Generating Ground-Truth Answers

**Question:** “Who is the mother of the father of the person whose hobby is reading?”

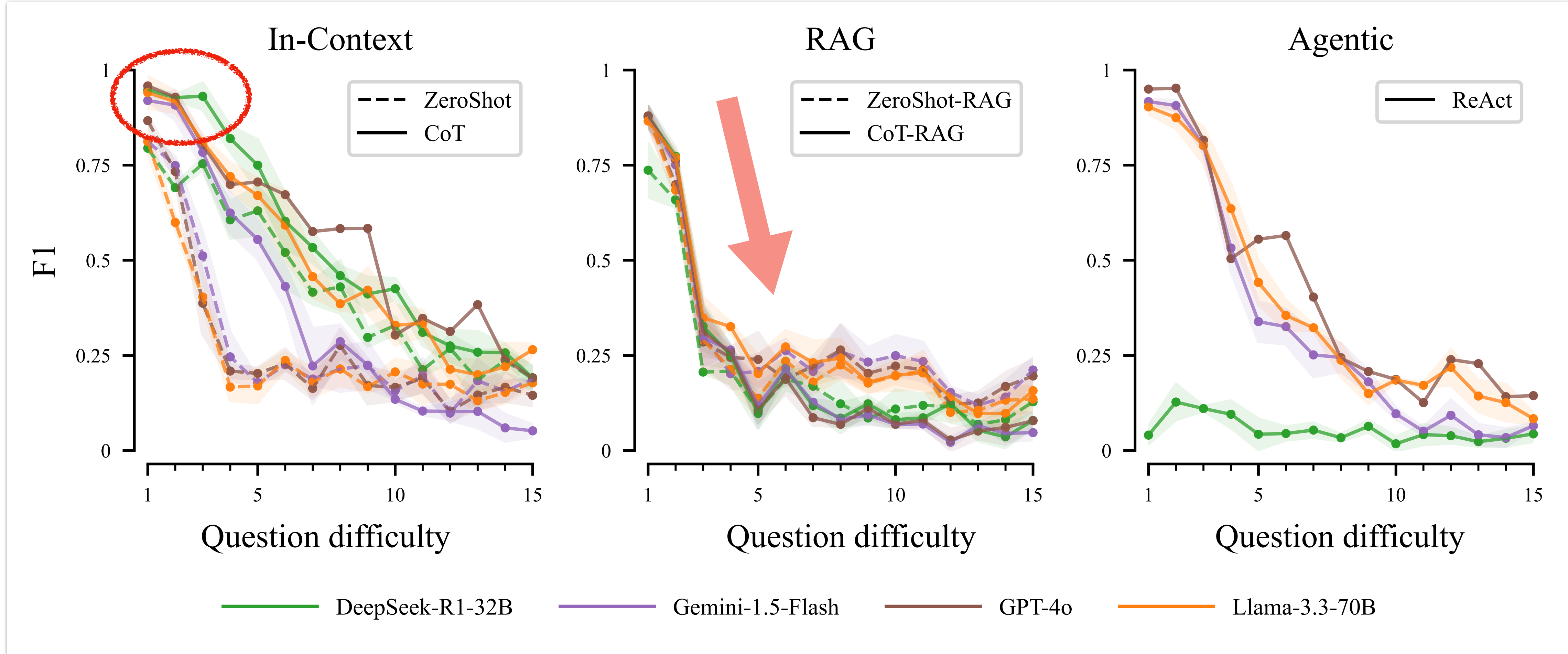
`hobby(X, “reading”), uncle(X, Y), mother(Y, Z)`



Prolog Database

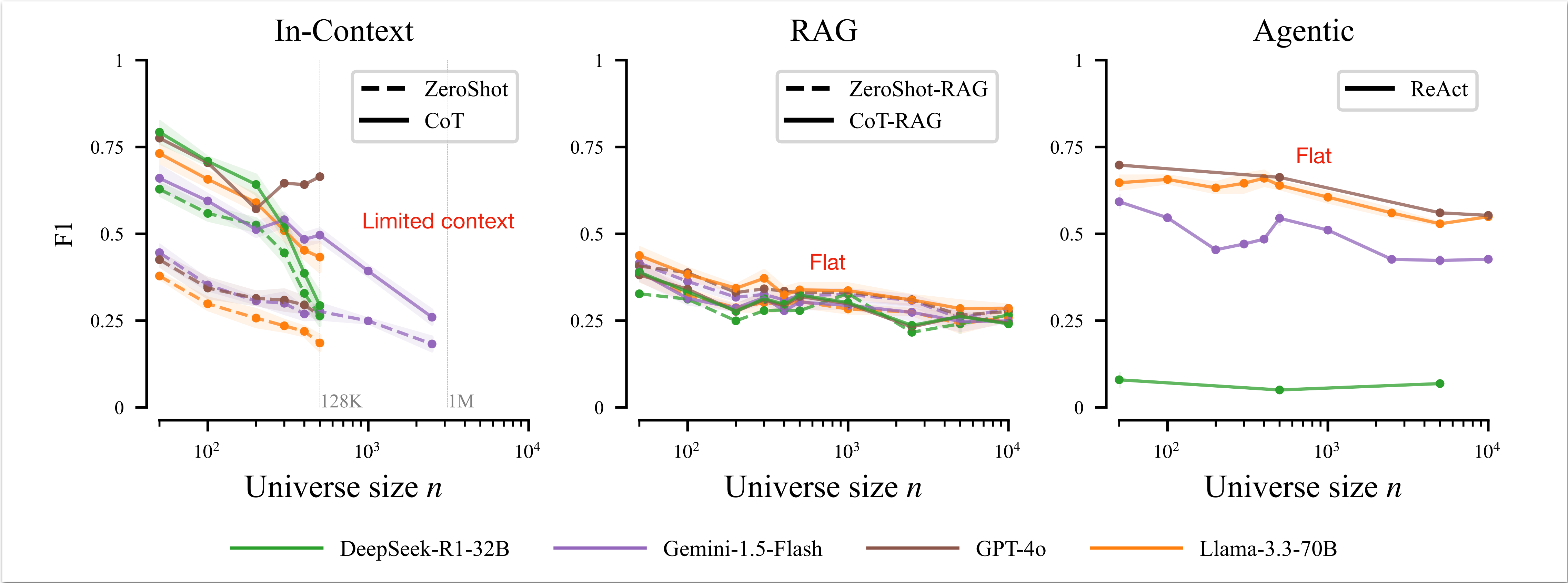
`Z←Cythia Smock`

# Evaluating Reasoning





# Evaluating Retrieval



# Qualitative Results

arxiv.org/abs/2506.20642

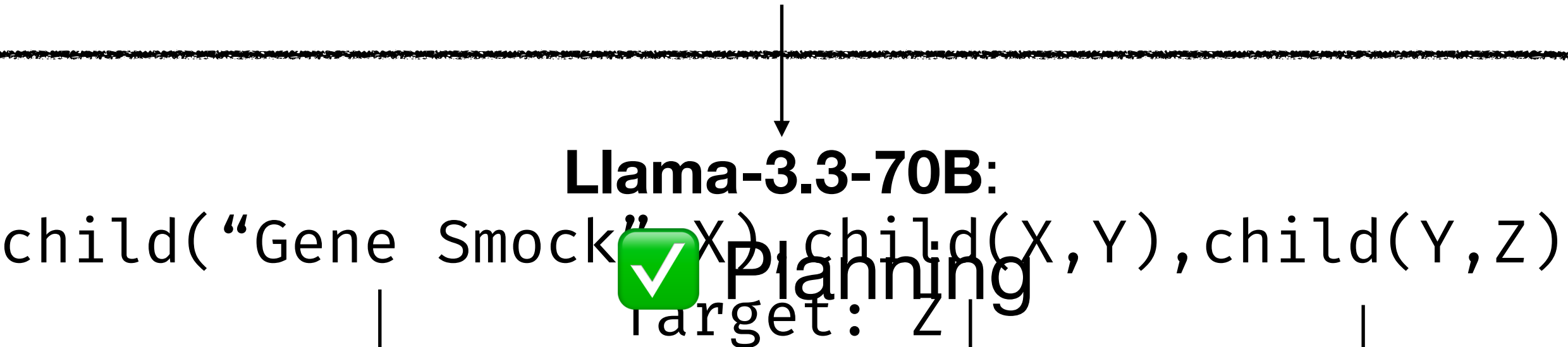


🌟 Memento 🌟

Who is the great-grandchild of Gene Smock?

Who is the great-grandchild of Gene Smock?

**Llama-3.3-70B:** To find the great-grandchild of Gene Smock, we first need to identify Gene Smock's children and then their children, and finally, their children.

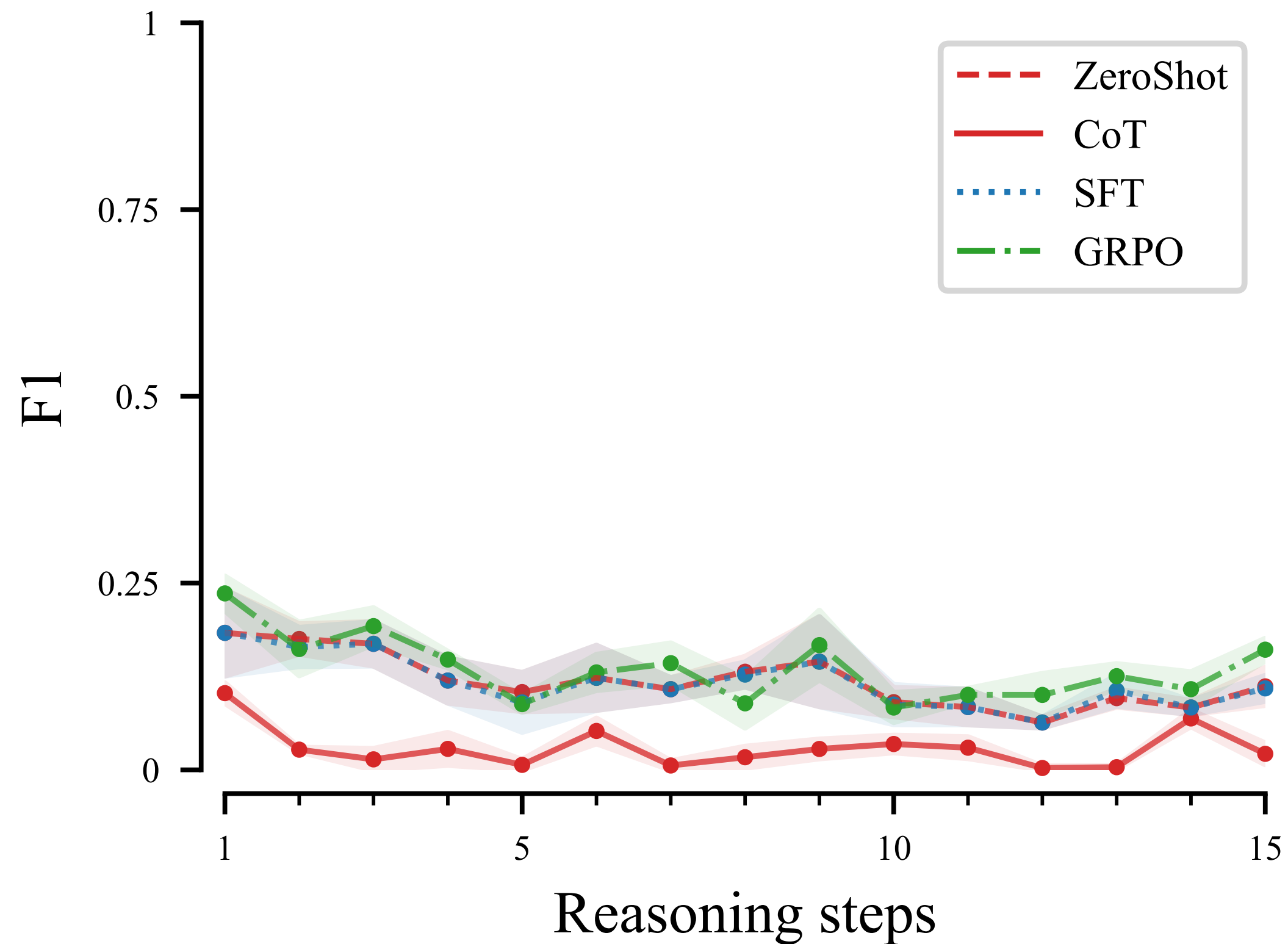


Gene Smock's child is Williams Smock.  
Williams Smock's children are Shelli Beltran and...  
Shelli Beltran's children are Aida Wang and...  
Aida Wang's children are Alfredo Wang...  
Then, we look at the children of Alfredo Wang...

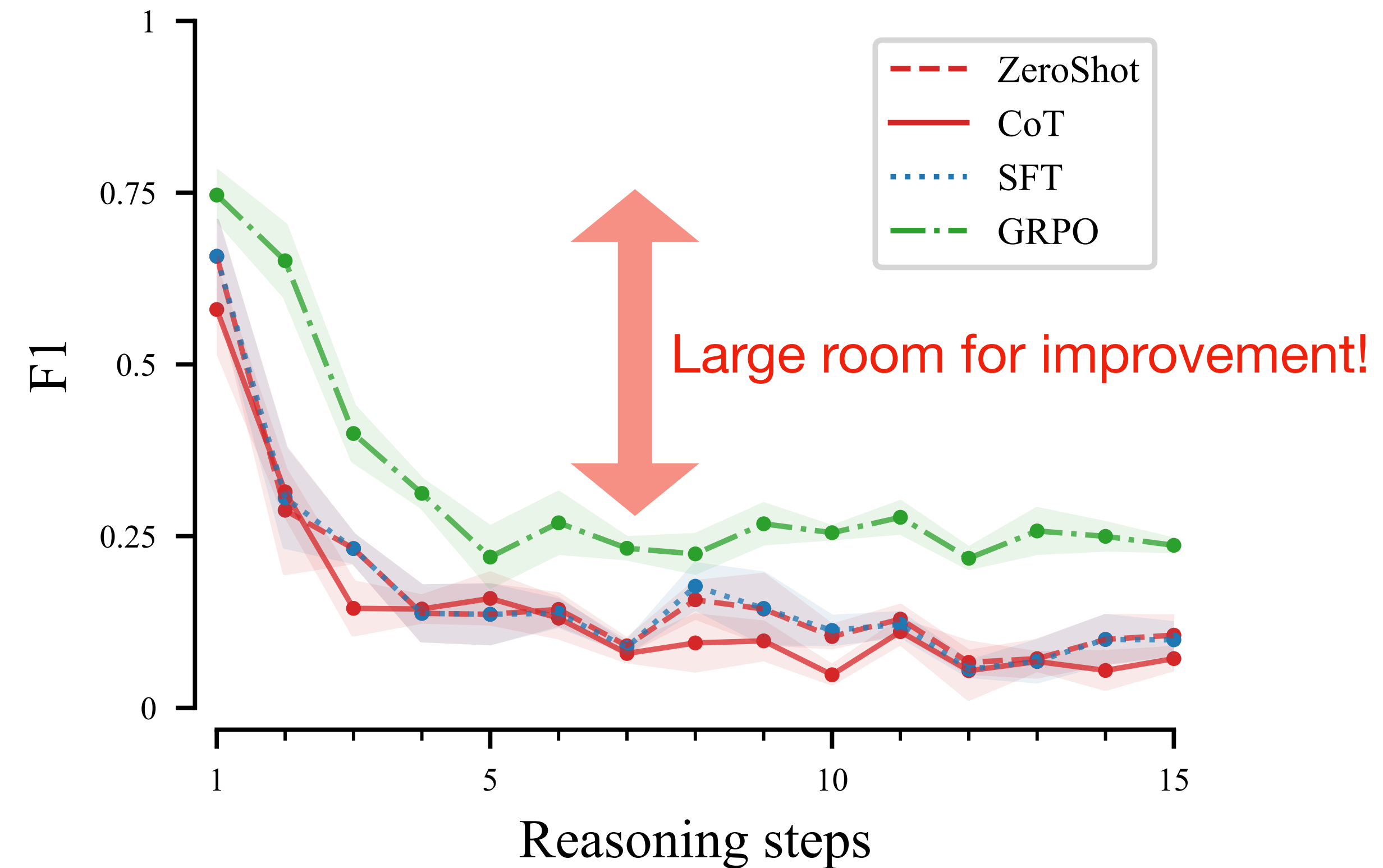


**HotpotQA:** 49.7→53.7  
**2WikiMultiHopQA:** 35.0→55.2  
**MuSiQue:** 15.7→21.9

# Training on PhantomWiki



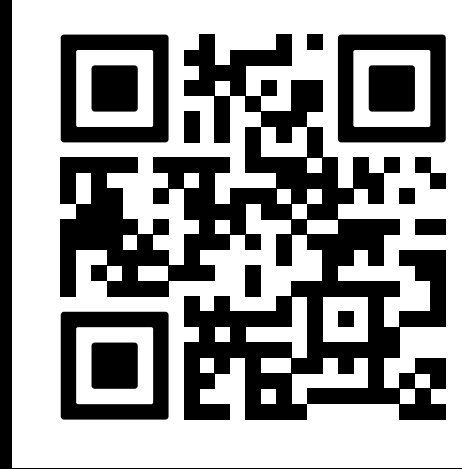
Qwen-2.5-0.5B-Instruct



Qwen-2.5-1.5B-Instruct (LoRA)



# Takeaways



[arxiv.org/abs/2502.20377](https://arxiv.org/abs/2502.20377)



[github.com/kilian-group/  
phantom-wiki](https://github.com/kilian-group/phantom-wiki)

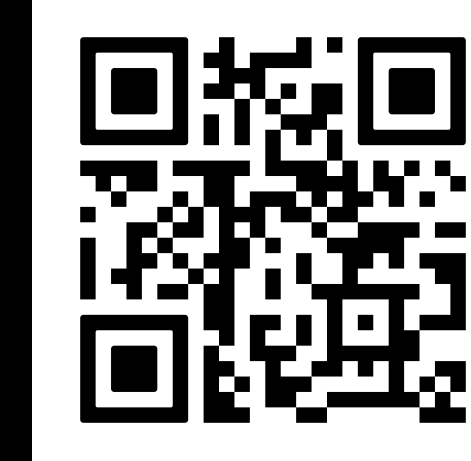
## Benefits of PhantomWiki:

- ✓ LLMs have no prior factual knowledge
- ✓ Factually consistent questions and answers
- ✓ Cheap (1M-sized dataset with 4-hop questions takes < 30 CPU hours!)
- 🤖 No LLMs (or humans) are harmed in the process! 🥰

# Future Directions



[arxiv.org/abs/2502.20377](https://arxiv.org/abs/2502.20377)



[github.com/kilian-group/  
phantom-wiki](https://github.com/kilian-group/phantom-wiki)

- How can we use Memento to bootstrap PhantomWiki?
- How to scale to 1 billion articles?
- How to utilize intermediate solutions from Prolog for training reasoning models?



**LMLM**  
✨ Pre-training Large  
Memory Language Models  
with Internal and External  
Knowledge ✨

