PhantomWiki: On-Demand Datasets for Reasoning & Retrieval Evaluation

Albert Gong*, Chao Wan*, Kamilė Stankevičiūtė*, Anmol Kabra*

Raphael Thesmar, Johann Lee, Julius Klenke, Carla P. Gomes, Kilian Q. Weinberger

*Equal contribution

UNIVERSITY OF CAMBRIDGE

arxiv.org/abs/2502.20377 github.com/kilian-group/phantom-wiki

Takeaways

PhantomWiki stress tests LLMs on multi-hop question-answering tasks, while ensuring:

1. LLMs cannot rely on factual knowledge from training
2. Factually consistent questions and answers
3. Fast and fully-automated dataset generation: 1M-sized dataset with 4-hop questions takes < 30 CPU hours!

Motivation

LLMs must answer questions with up-to-date knowledge, but evaluation is a challenge:

- Static datasets are prone to **data leakage**
- Disentangling a model's **internal knowledge, reasoning, and retrieval** capabilities is difficult with datasets curated from public data (e.g., Wikipedia).

Many Evaluation Scenarios

1. **In-Context:** document corpus fits in LLM context window [NIAH, LongBench, RULER, ∞-Bench, HELMET]
2. **RAG:** relevant documents are retrieved using an external retriever [MultiHop-RAG, BRIGHT, ARES]
3. **Agentic:** LLM uses external tools to obtain relevant context [ToolQA]

Prolog + Context-Free Grammar

Fact = predicate + constants

The mother of Alice is Charlotte.

mother("Alice", "Charlotte").

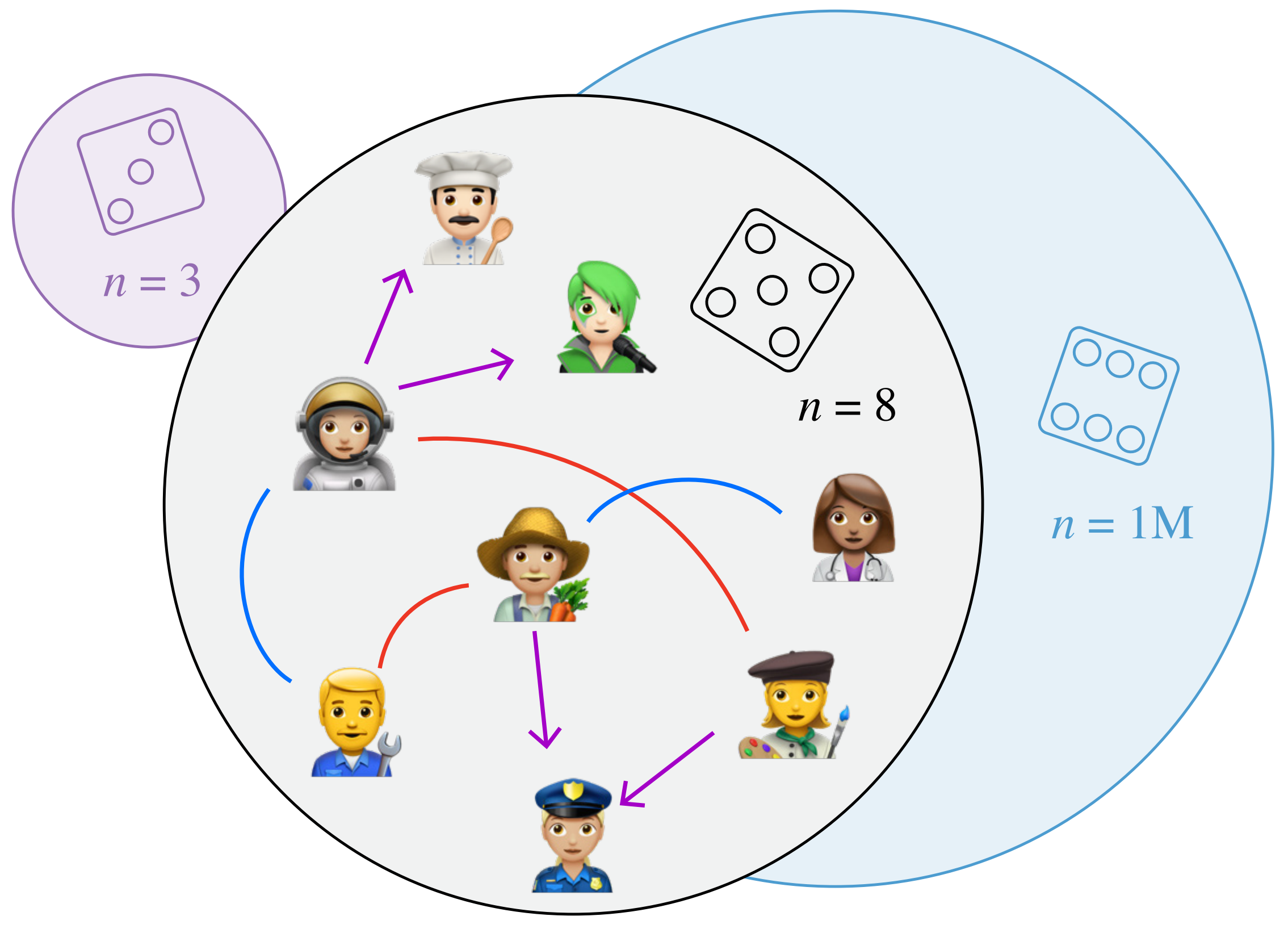
Query = predicate + variables

Who is the mother of Alice?

mother("Alice", X).

```
S  → Who is R ?
R  → the <relation> of
R’
R’ → R | <name>
```

PhantomWiki Pipeline



(1) Generate a random universe of size n

🧑‍🌾 **David Smith**

The friend of David is 👨‍🔧 John Harper.

The hobby of David is 🦅 birdwatching.

(2) Generate document corpus for the universe

```
Who is  <...>  ?
  the 🟦  of  <...>
    the 🟥  of  <...>
      the person whose 🟧 is 🟧

   🟦 → {nephew}  🟥 → {friend}
       🟧 🟧 → {hobby},{🦅}
```

(3) Generate questions using a context-free grammar

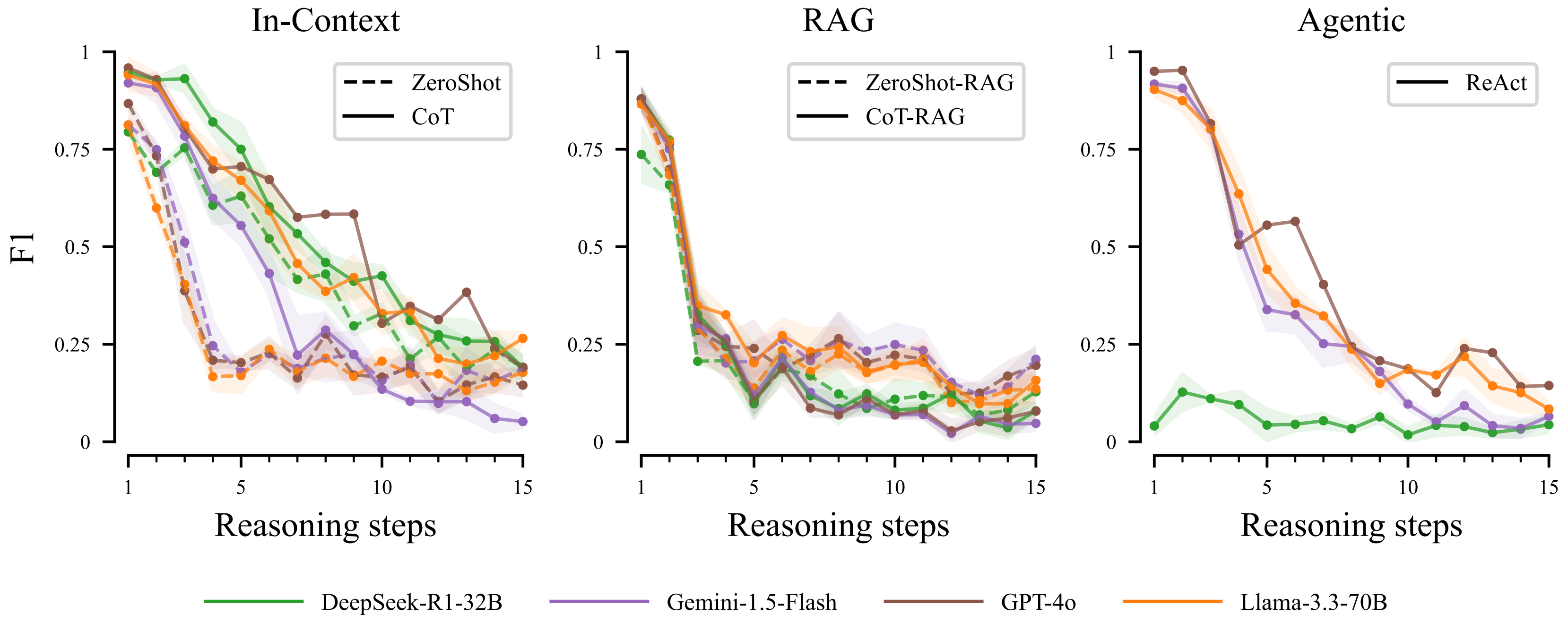
Q: Who is the nephew of the friend of the person whose hobby is birdwatching?

```
?- nephew(X2, Y),
   friend(X1, X2),
   hobby(X1, 🦅).
```

A: Y={👨‍🍳, 🧝}

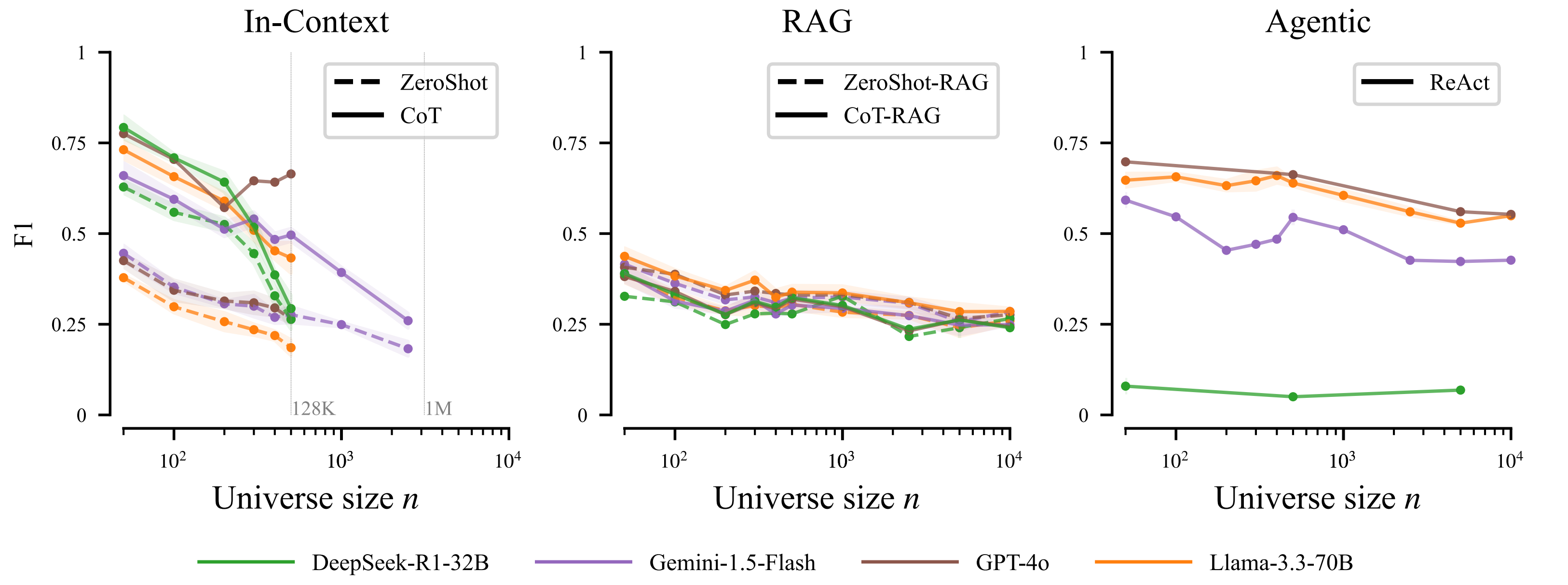
(4) Use Prolog to deduce ground-truth answers

PhantomWiki for reasoning evaluation

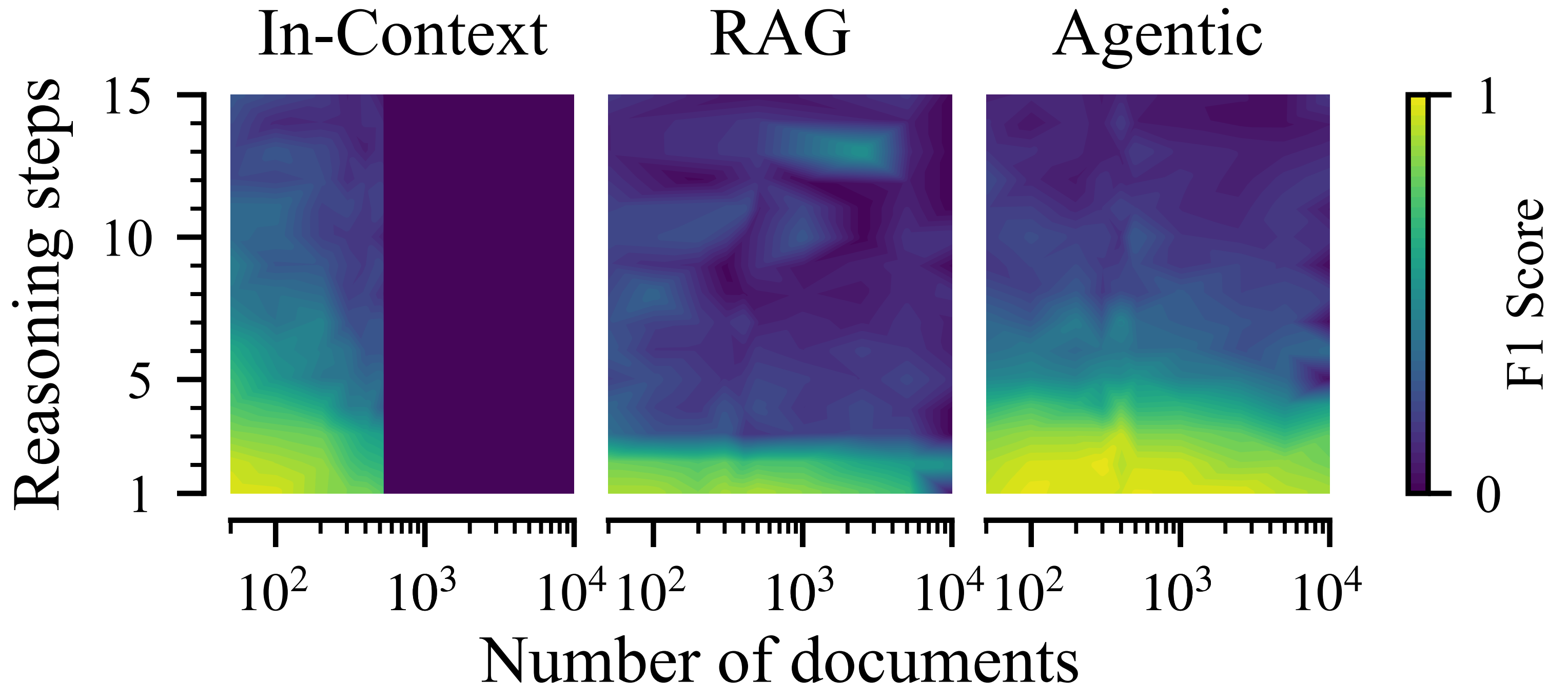


Reasoning steps = CFG depth + relation difficulty

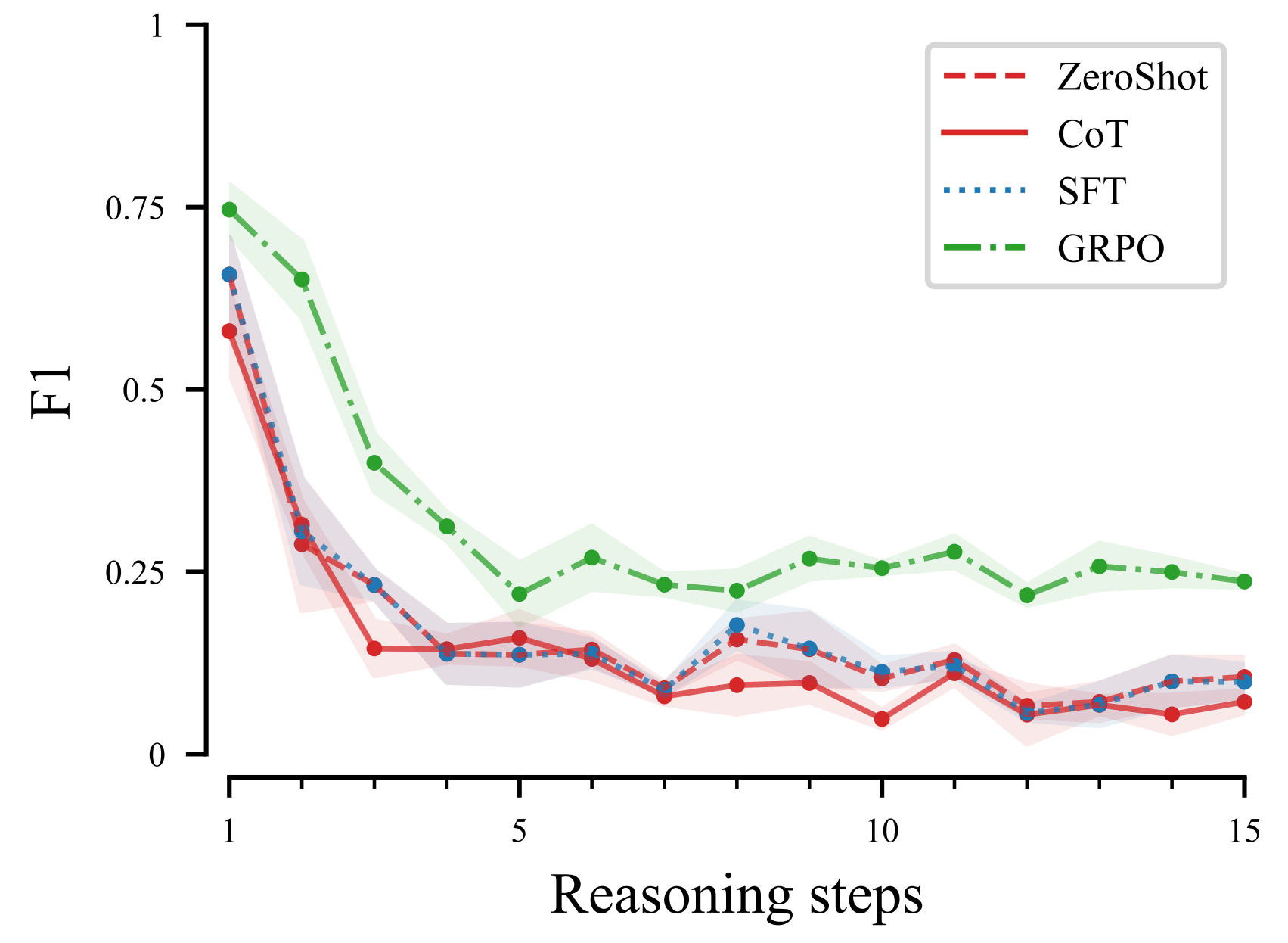
PhantomWiki for retrieval evaluation



Retrieval & Reasoning (w/ Llama 3.3-70B)

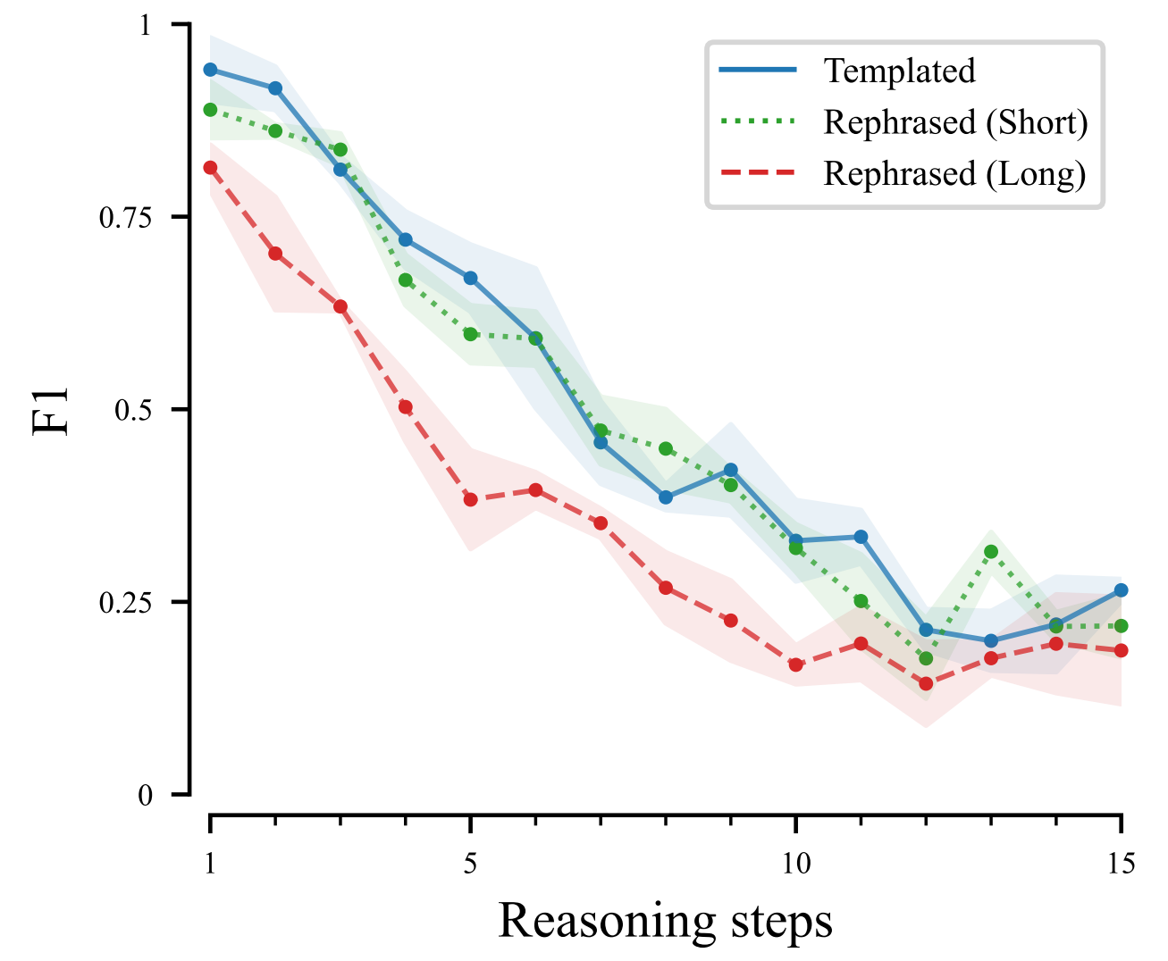


Does training improve reasoning ability?



LoRA on Qwen2.5-3B

Does article style affect reasoning ability?



Templated vs LLM-written