PhantomWiki: On-Demand Datasets for Reasoning and Retrieval Evaluation

Anonymous Authors[1]

Abstract

High-quality benchmarks are essential for evaluating reasoning and retrieval capabilities of large language models (LLMs). However, curating datasets for this purpose is not a permanent solution as they are prone to data leakage and inflated performance results. To address these challenges, we propose PhantomWiki: a pipeline to generate unique, factually consistent document corpora with diverse question-answer pairs. Unlike prior work, PhantomWiki is neither a fixed dataset, nor is it based on any existing data. Instead, a new PhantomWiki instance is generated on demand for each evaluation. We vary the question difficulty and corpus size to disentangle reasoning and retrieval capabilities respectively, and find that PhantomWiki datasets are surprisingly challenging for frontier LLMs. Thus, we contribute a scalable and data leakage-resistant framework for disentangled evaluation of reasoning, retrieval, and tool-use abilities.

1. Introduction

Designing agents that can perform complex reasoning while interfacing with a large-scale, dynamic corpus—like Wikipedia—is a long-standing goal in the field of natural language processing (Feldman & El-Yaniv, 2019; Min et al., 2019). Such a goal may be within reach given the impressive capabilities of recent language models, which are all trained on internet-scale data. For example, the ability of LLMs to solve math problems on GSM8K (Cobbe et al., 2021) and mathematical olympiads (AlphaProof & AlphaGeometry, 2024) could bode well for agents to answer highly quantitative questions. On benchmarks like DROP (Dua et al., 2019) and MMLU (Hendrycks et al., 2020), these LLMs demonstrate advanced reading comprehension and general reasoning capabilities, both necessary for intelligent agents. When augmented with retrievers (Muennighoff et al., 2022) and tools (Patil et al., 2023), LLMs seem to already possess a strong ability for accessing external datastores and knowledge bases.

However, it is unclear to what extent these models rely on their internal knowledge, which can easily become outdated, versus their reasoning and retrieval abilities. Consider the example, "*What is the date of birth of Wolfgang Amadeus Mozart?*". Since this fact is contained within LLMs' pretraining data, asking LLMs this question cannot provide reliable insight on whether the answer was deduced, retrieved or recalled. At the same time, existing approaches that perturb Wikipedia facts (Cohen et al., 2024; Meng et al., 2022; Elazar et al., 2021) to construct new question-answer pairs face challenges of ensuring factual consistency across articles. For example, changing Mozart's date of birth to 2025 also requires modifying Beethoven's article to erase the fact that Beethoven might have met Mozart in 1787!

One could hope to isolate reasoning from factual knowledge using mathematical or logical reasoning benchmarks. Unfortunately, such benchmarks are not entirely reliable as indicators of reasoning performance either. On GSM8K, a dataset of grade school math problems, Mirzadeh et al. (2024) report that frontier models perform significantly worse with minor or even meaningless alterations to the test data—indicating these models are vulnerable to overfitting at best and exact memorization at worst. To ensure fair comparison, LLMs need to be evaluated in a way that does not depend on any particular dataset instance.

Following this philosophy, we develop *PhantomWiki*. At the click of a button, PhantomWiki generates a fictional universe of characters along with a set of facts about them. We reflect these facts in a large-scale corpus, mimicking the style of fan-wiki websites. Then we generate question-answer pairs about the universe, encapsulating the types of multi-hop questions commonly considered in the question-answering (QA) literature. By adjusting the total context length—determined by PhantomWiki universe size—and the quantity of relevant information required for a given question, PhantomWiki provides a reliable benchmark for evaluating LLMs' in-context retrieval capabilities.

Our evaluation on PhantomWiki confirms that the proposed tasks present significant challenges for all of the state-of-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Under review by the Workshop on Long-Context Foundation Models (LCFM) at ICML 2025. Do not distribute.

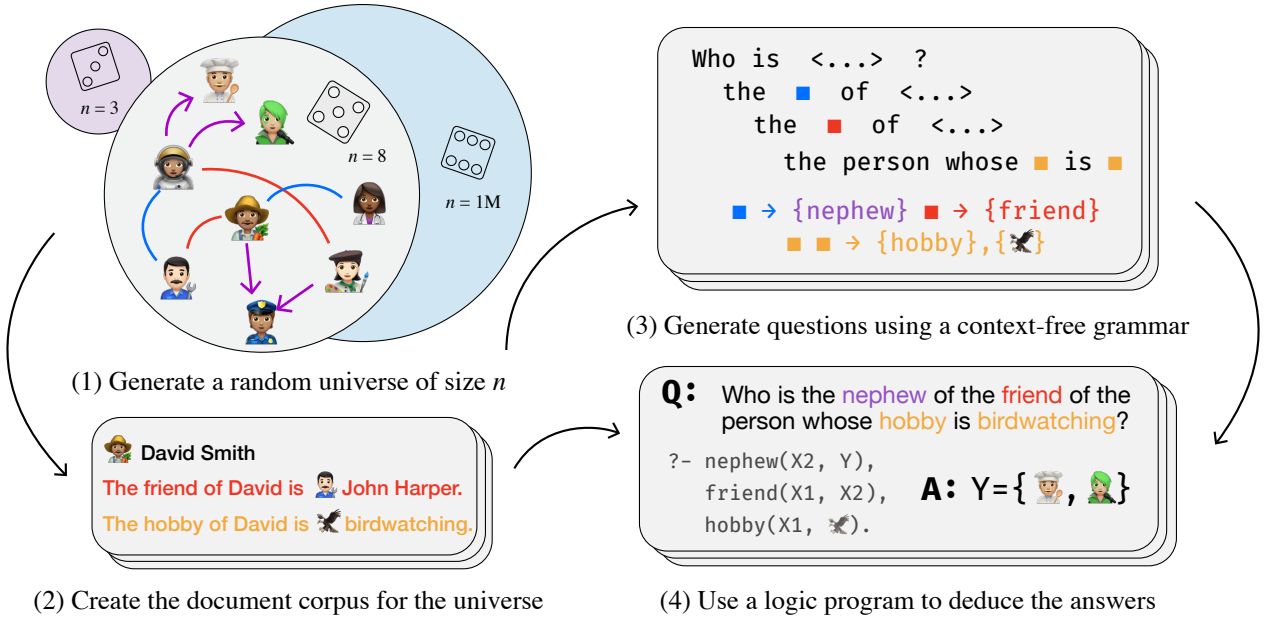


Figure 1. **Overview of the PhantomWiki pipeline.**

the-art LLMs that we used. Beyond serving as a robust benchmark for LLM performance, PhantomWiki provides valuable insights that can guide improvements in retrieval, reasoning, and tool-use capabilities of LLMs for the research community. We will make PhantomWiki code available in a public GitHub repository after the anonymous reviewing period.

2. PhantomWiki Construction

PhantomWiki is at its core an on-demand random generator of fictional worlds. Similarly to the wiki hosting services popular in film, video games, and literature[1], we represent these fictional worlds through Wikipedia-like biographical entries about their characters. We then test the model's understanding of the fictional world through an accompanying set of automatically generated question-answer pairs.

2.1. Generating a PhantomWiki Universe

The first stage of the PhantomWiki pipeline generates a random universe of n characters as well as the document corpus describing it, as illustrated in Figure 1, (1-2).

Generating Characters. Each character in a PhantomWiki universe is described through its *social relationships* and *personal facts* (Figure 1, (1)). For the social relationships, we first generate family trees, following the family tree generator of Hohenecker & Lukasiewicz (2020). We iteratively pick a person and generate their parent or child based on various constraints[2], until the user-specified universe size of n people is reached. The user can also specify other hyperparameters like the number of trees, their maximal depth, and the maximal number of offspring for each person. In addition to the family trees, we generate a friendship graph using the Erdős–Rényi model (making two people friends with some fixed probability, typically controlled by the desired average number of friendships.)

Generating Facts. Next, we generate personal facts for each person in the PhantomWiki universe. Names are assigned during the family generation procedure, with the first name sampled based on the character's gender and the surname based on the family tree, resulting in 15M full names in total[3]. We also add dates of birth in a way that is consistent with the existing family relations, and assign each person a job and a hobby that we uniformly sample from over 300 and 600 options respectively.

Generating Articles. Given all relevant facts for each person, we convert them into articles using pre-defined templates, e.g. "The job of David is a farmer. The hobby of David is birdwatching." (see Figure 1, (2)). This construction conveys the necessary information while keeping the articles short (about 160 tokens on average). While it is possible to extend the article generation process to LLM-based

[1]For example, see https://stardewvalley.fandom.com or https://harrypotter.fandom.com.

[2]For example, the number of offspring of a person has to be smaller than some threshold, parents of the people at the maximal tree level will not be generated, etc.

[3]We use unique names in our experiments, but PhantomWiki also supports repeated names.

methods (see e.g. Shao et al. 2024a), this poses the challenge of guaranteeing factual correctness without additional costs and external supervision. This has been supported by our preliminary experiments on article generation using Llama-3.3-70B, where we observed factual errors in the resulting articles; therefore we do not use LLMs and rely entirely on templates. The articles are the only component of PhantomWiki available to the model during its evaluation.

2.2. Generating Question-Answer Pairs

In the second half of the PhantomWiki pipeline, we generate a set of questions with verifiable answers, as shown in Figure 1, (3-4).

Generating Questions. We implement automatic question generation through a context-free grammar (CFG, Hopcroft et al. 2001) of *question templates*, which we then use to sample complete questions. For example, the question template "*Who is the* <relation> *of* <name>?" can be used to sample the question "*Who is the friend of David?*" (see Figure 1, (3)). The main advantage of using a CFG is that it efficiently and systematically obtains *all* possible compositions of questions for some recursion depth d. For instance, the following subset of our context-free grammar:

$$
\begin{aligned}
S &\to \text{Who is } R\,? \\
R &\to \text{the } \textit{<relation>} \text{ of } R' \\
R' &\to R \mid \textit{<name>}
\end{aligned}
$$

can lead to questions ranging from "*Who is the friend of David?*" to "*Who is the nephew of the friend of the brother of David?*" as d increases. In addition to these nested compositions, our CFG also supports questions about personal attributes (e.g. "*Who is the person whose hobby is birdwatching?*"), aggregation questions ("*How many brothers does David have?*"), and combinations of all three ("*How many friends does the brother of the person whose hobby is birdwatching have?*") (For the full CFG see Appendix B.)

Generating Answers. To ensure that the answers to the sampled questions are verifiably correct, we represent our generated universe in Prolog, a logic programming language (Sterling & Shapiro, 1994). Each Prolog program consists of a set of facts known about the world such as `hobby("David", "birdwatching")`, and a set of rules defining how facts are related to each other, such as `nephew(X, Y) :- sibling(X, A), son(A, Y)`. The Prolog program uses these facts and rules to deduce the exhaustive set of answers to its *queries* (i.e., the CFG-generated questions). For example, a question "*Who is the nephew of the friend of the person whose hobby is birdwatching?*" corresponds to the three-statement Prolog query `?- nephew(X2, Y), friend(X1, X2), hobby(X1, "birdwatching")`, which returns all people satisfying these constraints in the PhantomWiki universe (see Figure 1 (4)).

To construct the Prolog queries automatically, we modify the CFG algorithm to generate both the question and query templates in parallel. We note, however, that the queries are separate from the final PhantomWiki corpus and question-answer pairs, and the answers returned by the Prolog program should be held out as part of the evaluation procedure.

2.3. PhantomWiki Complexity

The goal of PhantomWiki is to generate memorization-resistant evaluation datasets that are challenging in both reasoning and retrieval aspects. In this section, we discuss our conceptual and practical design choices that help us achieve this goal.

Universe Space Complexity. To ensure that our evaluation with PhantomWiki is memorization and data leakage-resistant, we first show that the space of possible universes is sufficiently large to generate enough unique instances. Observe that the number of possible friendship assignments grows at the rate of $\Theta(2^{n^2})$ (Flajolet & Sedgewick, 2009, Ex. II.5) as the number of individuals n in the universe increases. Similarly, assuming each individual is assigned one fact from each category (job, hobby, etc.), the number of possible fact assignments grows at the rate $\Theta(c^n)$, where c is the total number of choices across the categories. PhantomWiki thus samples a corpus from $\Theta(2^{n^2}c^n)$ possible universes, which leads to diverse datasets optimal for data leakage-resistant evaluation. We note that as future work PhantomWiki could be extended to increase this diversity, e.g. by adding a temporal dimension of events.

Reasoning Complexity. The CFG enables us to recursively compose templates that lead to complex reasoning questions. Observe that our CFG in Appendix B produces $\Theta(d)$ question templates as the recursion depth d increases. Moreover, we can increase the difficulty of each template by increasing the number of *reasoning steps*. For example, substituting <*relation*> with *nephew* in a template adds two reasoning steps (`nephew(X, Y) :- sibling(X, A), son(A, Y)`), since PhantomWiki articles only contain immediate family relationships like *sibling* and *son*. In contrast, substituting <*relation*> with *second cousin* would lead to five reasoning steps. As we will show in Section 3, PhantomWiki questions are sufficiently complex to evaluate reasoning capabilities of state-of-the-art LLMs. We further note that PhantomWiki's CFG can be easily extended to support more question types like comparison and multiple-constraint questions.

3. Evaluating Reasoning

To isolate LLM reasoning capabilities with PhantomWiki, we investigate model performance on small universes ($n =$

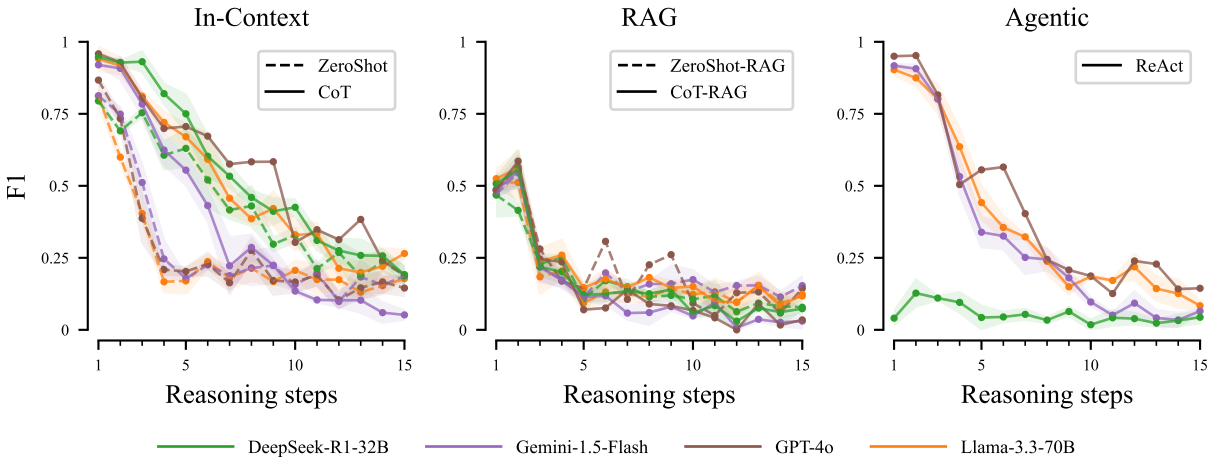


Figure 2. **F1 scores versus question difficulty, measured by *reasoning steps*.** We plot LLM performance on universe size $n = 50$, which is well within all models' context lengths, and report F1 scores averaged over 3 generation seeds. Increasing question difficulty in PhantomWiki reveals a clear decline across all state-of-the-art LLMs and prompting techniques, showing their struggle with reasoning.

50) in Figure 2. Note that contexts of all LLMs can fully include small universe document corpora. Each PhantomWiki dataset contains questions covering a wide range of difficulty. We evaluate three approaches: in-context prompting, RAG prompting, and agentic prompting. For each we plot the F1 scores as a function of question difficulty, as measured by the number of *reasoning steps* necessary to answer the question. As mentioned in Section 2.3, this is determined by the type of question templates and the sampled relationships. For all LLMs and prompting techniques, we verify empirically that **questions with larger reasoning steps are indeed more challenging to answer**. By allowing question difficulty to be adjusted, PhantomWiki serves as a foundational benchmark for evaluating reasoning capabilities in language models.

ZEROSHOT performance declines sharply as the number of reasoning steps increases for all LLMs, except for DeepSeek-R1-32B, which deteriorates more gradually. LLMs perform better with COT than with ZEROSHOT, but each additional reasoning step remains increasingly challenging. This suggests that even in the absence of retrieval constraints, LLMs struggle to navigate logical reasoning sequences.

RAG prompting techniques (ZEROSHOT-RAG and COT-RAG) stunt reasoning performance across the board—F1 scores are near zero on questions with 5 or more reasoning steps as opposed to 15 steps for in-context prompting. We attribute this to a core problem with RAG prompting: retrieving documents in the initial prompt before starting to answer the question, as opposed to reasoning through the question and retrieving documents dynamically.

We find that RAG prompting techniques can only answer questions that require a single reasoning step, like *Who is the friend of David?*. On the other hand, answering questions that require information from *multiple* reasoning steps is extremely challenging for ZEROSHOT-RAG and COT-RAG. To illustrate, consider the question *Who is the nephew of the friend of David?* Answering this question requires retrieving David's document first and then retrieving their friend's document to find the nephew. Since RAG prompting techniques retrieve documents *only once* by matching vector embeddings of questions and documents, they are unlikely to retrieve all necessary documents required to answer such questions.

Finally, the agentic prompting technique REACT allows LLMs to avoid the steep performance drop as seen in RAG prompting. On given a question, REACT prompting requires LLMs to retrieve documents dynamically in a conversation and justify why they are relevant. Concretely, before using a tool (`RetrieveArticle` or `Search`) in a conversation turn, the LLM is asked to describe how the tool will help using a "Thought" step (Yao et al., 2022), analogous to the COT prompting approach. This approach shows promise in answering questions correctly. Even so, REACT struggles as the question difficulty increases.

Figure 2 thus decomposes LLM performance along the lines of reasoning capabilities. It reveals that all in-context prompting and agentic prompting achieve near-perfect F1 scores on low-difficulty questions. Therefore, the performance of models only diverge on high-difficulty questions. To further isolate the impact of question difficulty, in Figure 5 we plot F1 scores as a function of reasoning steps for questions with only one solution.

Impact Statement

By leveraging context-free grammars and Prolog, PhantomWiki is able to generate large, durable and challenging datasets without using LLMs. The datasets have low computational, monetary, and environmental cost and our open-source framework is accessible to any user.

Since PhantomWiki randomly generates datasets that do not reference any existing data, the evaluation benchmark is resistant to data leakage and memorization while training. The approach of publishing a dataset generation *procedure* rather than a fixed dataset also encourages better research practices (by using fresh datasets instead of overfitting to a single instance), and enables a more accurate evaluation of model performance. Since we do not use any personal data, use of PhantomWiki does not have any privacy concerns.

References

AlphaProof & AlphaGeometry. AI achieves silver-medal standard solving International Mathematical Olympiad problems. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/, 2024. Accessed: 2025-01-25.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cohen, R., Biran, E., Yoran, O., Globerson, A., and Geva, M. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., and Goldberg, Y. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9: 1012–1031, 2021.

Feldman, Y. and El-Yaniv, R. Multi-hop paragraph retrieval for open-domain question answering. *arXiv preprint arXiv:1906.06606*, 2019.

Flajolet, P. and Sedgewick, R. *Analytic Combinatorics*. Cambridge University Press, 2009. ISBN 9781139477161. URL https://books.google.com/books?id=0h-4QcA1c1QC.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Hohenecker, P. and Lukasiewicz, T. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, 68:503–540, 2020.

Hopcroft, J. E., Motwani, R., and Ullman, J. D. Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32(1):60–65, 2001.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Min, S., Chen, D., Hajishirzi, H., and Zettlemoyer, L. A discrete hard em approach for weakly supervised question answering. *arXiv preprint arXiv:1909.04849*, 2019.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023.

Shao, Y., Jiang, Y., Kanell, T. A., Xu, P., Khattab, O., and Lam, M. S. Assisting in writing Wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*, 2024a.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Sterling, L. and Shapiro, E. Y. *The art of Prolog: advanced programming techniques*. MIT press, 1994.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

A. Background

A.1. Context-Free Grammars

Context-free grammar (CFG) is a type of formal grammar where the productions rules govern how to generate text from non-terminals and terminals. A context-free grammar is defined by $G = (V, \Sigma, R, S)$ where V and Σ denotes nonterminal and terminal respectively. R is a finite relation in $V \times (V \cup \Sigma)^*$ which specifies the production rules of the grammar. $S \in V$ is the start symbol. A production rule in R has the form

$$\alpha \to \beta \tag{1}$$

where $\alpha \in V$, $\beta \in (V \cup \Sigma)^*$. It is conventional to list all rules with the same left-hand side on the same line and separate the right-hand side with "|" like $\alpha \to \beta_1 \mid \beta_2$.

B. Question Template Generation

B.1. Context-Free Grammar

We use the following CFG to generate question templates:

```
S -> Who is R? | What is A ? | How many RN_p does R_c have ?
R -> the RN of R_c | the person whose AN is AV
R_c -> R | N
A -> the AN of R
RN -> <relation>
RN_p -> <relation_plural>
AN -> <attribute_name>
AV -> <attribute_value>
N -> <name>
```

B.2. CFG-generated question templates

Our CFG produces the following 50 question templates at recursion depth $d = 20$. Note how the recursive production rule `R_c -> R | N` leads to chained productions.

1. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of the <relation>_13 of the <relation>_15 of the <relation>_17 of the person whose <attribute_name>_19 is <attribute_value>_19?
2. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of the <relation>_13 of the <relation>_15 of the <relation>_17 of <name>_18?
3. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of the <relation>_13 of the <relation>_15 of the person whose <attribute_name>_17 is <attribute_value>_17?
4. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of the <relation>_13 of the <relation>_15 of <name>_16?
5. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of the <relation>_13 of the person whose <attribute_name>_15 is <attribute_value>_15?
6. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of the <relation>_13 of <name>_14?
7. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of the person whose <attribute_name>_13 is <attribute_value>_13?
8. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the <relation>_11 of <name>_12?

9. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of the person whose <attribute_name>_11 is <attribute_value>_11?
10. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the <relation>_9 of <name>_10?
11. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of the person whose <attribute_name>_9 is <attribute_value>_9?
12. Who is the <relation>_3 of the <relation>_5 of the <relation>_7 of <name>_8?
13. Who is the <relation>_3 of the <relation>_5 of the person whose <attribute_name>_7 is <attribute_value>_7?
14. Who is the <relation>_3 of the <relation>_5 of <name>_6?
15. Who is the <relation>_3 of the person whose <attribute_name>_5 is <attribute_value>_5?
16. Who is the <relation>_3 of <name>_4?
17. Who is the person whose <attribute_name>_3 is <attribute_value>_3?
18. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of the <relation>_18 of <name>_19?
19. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of the person whose <attribute_name>_18 is <attribute_value>_18?
20. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of <name>_17?
21. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the person whose <attribute_name>_16 is <attribute_value>_16?
22. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of <name>_15?
23. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the person whose <attribute_name>_14 is <attribute_value>_14?
24. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of <name>_13?
25. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the person whose <attribute_name>_12 is <attribute_value>_12?
26. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of <name>_11?
27. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of the person whose <attribute_name>_10 is <attribute_value>_10?
28. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the <relation>_8 of <name>_9?
29. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the person whose <attribute_name>_8 is <attribute_value>_8?
30. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of <name>_7?
31. What is the <attribute_name>_3 of the <relation>_4 of the person whose <attribute_name>_6 is <attribute_value>_6?
32. What is the <attribute_name>_3 of the <relation>_4 of <name>_5?
33. What is the <attribute_name>_3 of the person whose <attribute_name>_4 is <attribute_value>_4?

34. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of the <relation>_18 of <name>_19 have?
35. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of the person whose <attribute_name>_18 is <attribute_value>_18 have?
36. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of <name>_17 have?
37. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the person whose <attribute_name>_16 is <attribute_value>_16 have?
38. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of <name>_15 have?
39. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the person whose <attribute_name>_14 is <attribute_value>_14 have?
40. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of <name>_13 have?
41. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the person whose <attribute_name>_12 is <attribute_value>_12 have?
42. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of <name>_11 have?
43. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the person whose <attribute_name>_10 is <attribute_value>_10 have?
44. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of <name>_9 have?
45. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the person whose <attribute_name>_8 is <attribute_value>_8 have?
46. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of <name>_7 have?
47. How many <relation_plural>_2 does the <relation>_4 of the person whose <attribute_name>_6 is <attribute_value>_6 have?
48. How many <relation_plural>_2 does the <relation>_4 of <name>_5 have?
49. How many <relation_plural>_2 does the person whose <attribute_name>_4 is <attribute_value>_4 have?
50. How many <relation_plural>_2 does <name>_3 have?

B.3. Question-Answer Characteristics

C. Baseline Details

C.1. LLM Sampling Hyperparameters

	Temperature	Top-k	Top-p	Repetition Penalty	Sampling Seed	Max number of output Tokens
Values	0	50	0.7	1.0	0	4096

Table 1. Default Hyperparameters values for LLM Sampling

We used the above default hyperparameters values for all models, but DeepSeek-R1-32B, where we used temperature $= 0.6$ and top-p $= 0.95$.

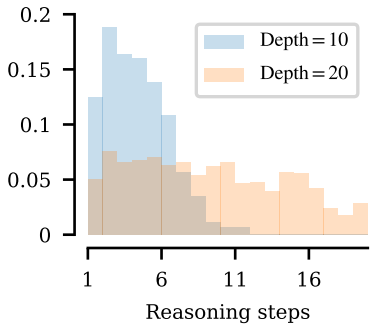


Figure 3. Histogram of question difficulties (measured by reasoning steps) for universe size $n = 50$ at two CFG recursion depths $d \in \{10, 20\}$. We average the frequencies across 3 dataset generation seeds.

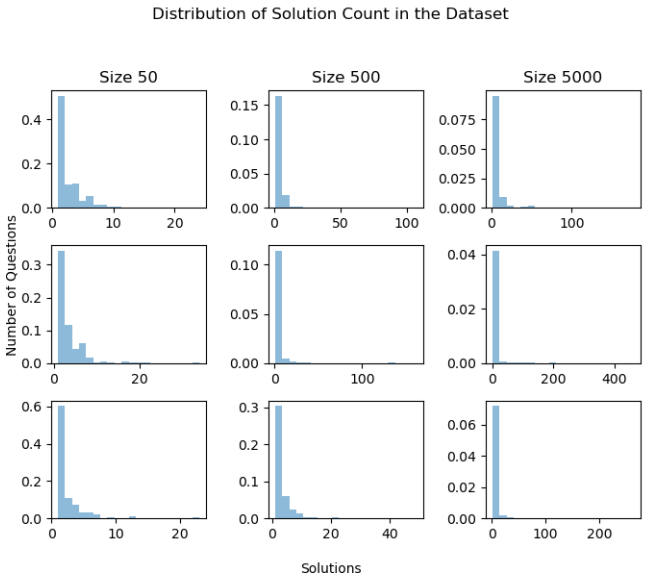


Figure 4. Distribution of number of answers across sizes $n \in \{50, 500, 5000\}$, seeds $\{1, 2, 3\}$, and CFG depth 20.

C.2. ZEROSHOT-SIMPLE

We use the following prompt for all models, where `evidence` is the concatenation of all documents in the PhantomWiki instance.

```
You are given the following evidence:
(BEGIN EVIDENCE)
{{evidence}}
(END EVIDENCE)

You will be provided a question. Your task is to provide an answer according to
    these instructions:
- The output must be one of the following: a name (if there is only one correct
    answer); or a list of names separated by '{constants.answer_sep}' (if there
    are multiple correct answers).
- DO NOT include any additional information in your answer.

Question: {{question}}
Answer:
```

For DeepSeek-R1-32B, we additionally parse the output to separate the model's reasoning process from its final answer using the `</think>` tag.

C.3. ZEROSHOT-RAG

The prompt is exactly the same as ZEROSHOT, except we replace `evidence` with 4 documents retrieved using the UAE-LARGE-V1. We pre-compute an index the document corpus using an FAISS vector store of UAE-LARGE-V1 embeddings. Upon generation, we search for similar documents for `question` according to maximum inner product search on document and question embeddings.

C.4. CHAIN-OF-THOUGHT-SIMPLE

We use the following prompt for all models, where `evidence` is replaced with a list of all documents. We use a regular expression to parse the output.

```
You are given the following evidence:
(BEGIN EVIDENCE)
{{evidence}}
(END EVIDENCE)

You will be provided a question. Your response must end in the following sentence
    : The answer is <answer>.
Here, <answer> must be one of the following:
- a name (if there is only one correct answer); or
- a list of names separated by '{constants.answer_sep}' (if there are multiple
    correct answers).

Here are some examples:
(START OF EXAMPLES)
Example 1:
Question: Who is the brother of Dino Beltran?
Answer: Based on the evidence, the brother of Dino Beltran is Orlando Beltran.
    The answer is Orlando Beltran.

Example 2:
Question: Who is the sibling of Barabara Beltran?
Answer: Based on the evidence, the siblings of Barabara Beltran are Aida Wang,
    Vicki Hackworth. The answer is Aida Wang{constants.answer_sep}Vicki Hackworth
    .

Example 3:
Question: Who is the child of the sibling of Stacia Toombs?
Answer: First I need to find the sibling of Stacia Toombs. Based on the evidence,
     the sibling of Stacia Toombs is Shelli Beltran. Now I need to find the child
     of Shelli Beltran. Based on the evidence, the children of Shelli Beltran are
     Aida Wang, Barabara Beltran, Vicki Hackworth. The answer is Aida Wang{
    constants.answer_sep}Barabara Beltran{constants.answer_sep}Vicki Hackworth.

Example 4:
Question: Who is the uncle of William Smock?
Answer: An uncle is the brother of a parent. Based on the evidence, the parents
    of William Smock are Dominique Smock, Gene Smock. To find the uncle of
    William Smock, I need to find the brother of Dominique Smock and Gene Smock.
    Based on the evidence, Dominique Smock has no brother, and the brother of
    Gene Smock is Eli Smock. So the uncle of William Smock is Eli Smock. The
```

```
   answer is Eli Smock.

Example 5:
Question: What is the occupation of the sister of the grandmother of Virgil
   Hackworth?
Answer: A grandmother is the mother of a parent. Based on the evidence, the
   parents of Virgil Hackworth are Ricardo Hackworth, Vicki Hackworth. To find
   the grandmother of Virgil Hackworth, I need to find the mother of Ricardo
   Hackworth and Vicki Hackworth. Based on the evidence, Ricardo Hackworth has
   no mother, and the mother of Vicki Hackworth is Shelli Beltran. Now I need to
    find the sister of Shelli Beltran. Based on the evidence, the sister of
   Shelli Beltran is Stacia Toombs. Based on the evidence, the occupation of
   Stacia Toombs is actuary. The answer is actuary.

Example 6:
Question: Who is the brother of the person whose occupation is associate
   professor?
Answer: I need to search for people whose occupation is associate professor.
   Based on the evidence, the person whose occupation is associate professor is
   Dino Beltran. And the brother of Dino Beltran is Orlando Beltran. The answer
   is Orlando Beltran.

Example 7:
Question: What is the date of birth of the person whose hobby is meteorology?
Answer: I need to search for people whose hobby is meteorology. Based on the
   evidence, the people whose hobby is meteorology are Alison Smock, Barabara
   Beltran. The date of birth of Alison Smock is 0929-10-28, and the date of
   birth of Barabara Beltran is 0989-06-11. The answer is 0929-10-28{constants.
   answer_sep}0989-06-11.

Example 8:
Question: Who is the cousin of the person whose occupation is broadcast engineer?
Answer: I need to search for people whose occupation is broadcast engineer. Based
    on the evidence, the person whose occupation is broadcast engineer is
   Barabara Beltran. A cousin is the child of the sibling of the parent. Based
   on the evidence, the parents of Barabara Beltran are Dino Beltran, Shelli
   Beltran. The sibling of Dino Beltran is Orlando Beltran, and the sibling of
   Shelli Beltran is Stacia Toombs. Based on the evidence, Orlando Beltran has
   no child, and the child of Stacia Toombs is Leslee Toombs. So the cousin of
   Barabara Beltran is Leslee Toombs. The answer is Leslee Toombs.

Example 9:
Question: Who is the great-granddaughter of the person whose hobby is biology?
Answer: I need to search for people whose hobby is biology. Based on the evidence
   , the person whose hobby is biology is Alvaro Smock. To find the great-
   granddaughter of Alvaro Smock, I need to find the daughter of the child of
   the child of Alvaro Smock. Based on the evidence, the children of Alvaro
   Smock are Eli Smock, Gene Smock. Eli Smock has no child, and the child of
   Gene Smock is Williams Smock. The daughters of Williams Smock are Shelli
   Beltran, Stacia Toombs. So the great-granddaughters of Alvaro Smock, whose
   hobby is biology, are Shelli Beltran, Stacia Toombs. The answer is Shelli
   Beltran{constants.answer_sep}Stacia Toombs.
(END OF EXAMPLES)
```

```
Question: {{question}}
Answer:
```

C.5. CHAIN-OF-THOUGHT-RAG

The prompt is exactly the same as CoT, except we replace `evidence` with 4 documents retrieved using the UAE-LARGE-V1. We use the same retriever setup as described in Appendix C.3.

C.6. REACT

We modify the ReAct agent implementation of Shinn et al. (2024), which can be found at https://github.com/noahshinn/reflexion, to support our keyword-based search tools `RetrieveArticle` and `Search`. As seen in the following examples, the `RetrieveArticle` tool fetches the documents by title, and the `Search` tool fetches all document titles whose bodies contain the argument. We instruct the LLM to output predict the answer using a third tool `Finish`.

```
Solve a question answering task with interleaving Thought, Action, Observation
    steps.
Thought can reason about the current situation, and Action can be 3 types:
(1) RetrieveArticle[{{{{entity}}}}]. This action retrieves the article about {{{{
    entity}}}} if it exists. If the article does not exist, the action will say
    so.
(2) Search[{{{{attribute}}}}]. This action searches the database for {{{{
    attribute}}}} and retrieves all articles that contain {{{{attribute}}}}. If
    no article contains {{{{attribute}}}}, the action will say so.
(3) Finish[{{{{answer}}}}]. This action answers the question with {{{{answer}}}}.
If you cannot find the answer, output the empty answer like: Finish[].
If there are multiple answers A,B,C, answer with a list like: Finish[A{constants.
    answer_sep}B{constants.answer_sep}C].

You may take as many steps as necessary.
Here are some examples:
(START OF EXAMPLES)
Example 1:
Question: Who is the brother of Dino Beltran?
Thought 1: I need to retrieve article about Dino Beltran and find who his brother
     is.
Action 1: RetrieveArticle[Dino Beltran].
Observation 1: # Dino Beltran ## Family The parents of Dino Beltran are Brian
    Beltran, Daisy Beltran.  Dino Beltran's sibling is Orlando Beltran.  The
    brother of Dino Beltran is Orlando Beltran.  The mother of Dino Beltran is
    Daisy Beltran.  The father of Dino Beltran is Brian Beltran.  The children of
     Dino Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth.  The
    daughters of Dino Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth.
    The wife of Dino Beltran is Shelli Beltran.  ## Friends The friend of Dino
    Beltran is Alvaro Smock.  ## Attributes The date of birth of Dino Beltran is
    0958-08-09.  The occupation of Dino Beltran is associate professor.  The
    hobby of Dino Beltran is shogi.
Thought 2: The brother of Dino Beltran is Orlando Beltran, so the answer is
    Orlando Beltran.
Action 2: Finish[Orlando Beltran].

Example 2:
Question: Who is the mother of Ivana Smith?
```

```
Thought 1: I need to retrieve article about Ivana Smith and find who her mother
   is.
Action 1: RetrieveArticle[Ivana Smith].
Observation 1: No article exists for the requested entity. Please try retrieving
   article for another entity.
Thought 2: Article about Ivana Smith does not exist, so I cannot find the mother
   of Ivana Smith. I will output empty answer.
Action 2: Finish[].

Example 3:
Question: Who is the sibling of Barabara Beltran?
Thought 1: I need to retrieve article about Barabara Beltran and find who her
   sibling is.
Action 1: RetrieveArticle[Barabara Beltran].
Observation 1: # Barabara Beltran ## Family The parents of Barabara Beltran are
   Dino Beltran, Shelli Beltran.  Barabara Beltran's siblings are Aida Wang,
   Vicki Hackworth.  The sisters of Barabara Beltran are Aida Wang, Vicki
   Hackworth.  The mother of Barabara Beltran is Shelli Beltran.  The father of
   Barabara Beltran is Dino Beltran.  ## Friends ## Attributes The date of birth
    of Barabara Beltran is 0989-06-11.  The occupation of Barabara Beltran is
   broadcast engineer.  The hobby of Barabara Beltran is meteorology.
Thought 2: The siblings of Barabara Beltran are Aida Wang, Vicki Hackworth. So
   the answer is Aida Wang, Vicki Hackworth.
Action 2: Finish[Aida Wang{constants.answer_sep}Vicki Hackworth].

Example 4:
Question: Who is the child of the sibling of Stacia Toombs?
Thought 1: First, I need to retrieve article about Stacia Toombs and find who her
    sibling is.
Action 1: RetrieveArticle[Stacia Toombs].
Observation 1: # Stacia Toombs ## Family The parents of Stacia Toombs are Alison
   Smock, Williams Smock.  Stacia Toombs's sibling is Shelli Beltran.  The
   sister of Stacia Toombs is Shelli Beltran.  The mother of Stacia Toombs is
   Alison Smock.  The father of Stacia Toombs is Williams Smock.  The child of
   Stacia Toombs is Leslee Toombs.  The daughter of Stacia Toombs is Leslee
   Toombs.  The husband of Stacia Toombs is Wilbert Toombs.  ## Friends The
   friends of Stacia Toombs are Brian Beltran, Isiah Lutz, Leeann Hackworth,
   Lesley Lutz, Ryan Wang.  ## Attributes The date of birth of Stacia Toombs is
   0959-03-22.  The occupation of Stacia Toombs is actuary.  The hobby of Stacia
    Toombs is finance.
Thought 2: The sibling of Stacia Toombs is Shelli Beltran. Now I need to retrieve
    article about Shelli Beltran and find who her child is.
Action 2: RetrieveArticle[Shelli Beltran].
Observation 2: # Shelli Beltran ## Family The parents of Shelli Beltran are
   Alison Smock, Williams Smock.  Shelli Beltran's sibling is Stacia Toombs.
   The sister of Shelli Beltran is Stacia Toombs.  The mother of Shelli Beltran
   is Alison Smock.  The father of Shelli Beltran is Williams Smock.  The
   children of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth.
    The daughters of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki
   Hackworth.  The husband of Shelli Beltran is Dino Beltran.  ## Friends The
   friends of Shelli Beltran are Brian Beltran, Eli Smock, Isiah Lutz, Leslee
   Toombs, Lesley Lutz, Ryan Wang.  ## Attributes The date of birth of Shelli
   Beltran is 0958-03-08.  The occupation of Shelli Beltran is occupational
   therapist.  The hobby of Shelli Beltran is sociology.
```

Thought 3: The child of Shelli Beltran is Aida Wang, Barabara Beltran, Vicki Hackworth. So the answer is Aida Wang, Barabara Beltran, Vicki Hackworth.
Action 3: Finish[Aida Wang{constants.answer_sep}Barabara Beltran{constants.answer_sep}Vicki Hackworth].

Example 5:
Question: Who is the uncle of William Smock?
Thought 1: First, I need to retrieve article about William Smock. To find the uncle of William Smock, I need to find the brother of his parent.
Action 1: RetrieveArticle[William Smock].
Observation 1: # Williams Smock ## Family The parents of Williams Smock are Dominique Smock, Gene Smock. The mother of Williams Smock is Dominique Smock. The father of Williams Smock is Gene Smock. The children of Williams Smock are Shelli Beltran, Stacia Toombs. The daughters of Williams Smock are Shelli Beltran, Stacia Toombs. The wife of Williams Smock is Alison Smock. ## Friends The friend of Williams Smock is Lannie Smock. ## Attributes The date of birth of Williams Smock is 0926-04-04. The occupation of Williams Smock is clinical biochemist. The hobby of Williams Smock is social studies.
Thought 2: The parents of William Smock are Dominique Smock, Gene Smock. To find the uncle of William Smock, I need to find the brother of Dominique Smock and Gene Smock.
Action 2: RetrieveArticle[Dominique Smock].
Observation 2: # Dominique Smock ## Family The child of Dominique Smock is Williams Smock. The son of Dominique Smock is Williams Smock. The husband of Dominique Smock is Gene Smock. ## Friends The friends of Dominique Smock are Gene Smock, Isiah Lutz, Orlando Beltran, Vicki Hackworth. ## Attributes The date of birth of Dominique Smock is 0897-09-08. The occupation of Dominique Smock is sports therapist. The hobby of Dominique Smock is dominoes.
Thought 3: Dominique Smock has no brother. Now I need to retrieve article about Gene Smock and find his brother.
Action 3: RetrieveArticle[Gene Smock].
Observation 3: # Gene Smock ## Family The parents of Gene Smock are Alvaro Smock, Lannie Smock. Gene Smock's sibling is Eli Smock. The brother of Gene Smock is Eli Smock. The mother of Gene Smock is Lannie Smock. The father of Gene Smock is Alvaro Smock. The child of Gene Smock is Williams Smock. The son of Gene Smock is Williams Smock. The wife of Gene Smock is Dominique Smock. ## Friends The friends of Gene Smock are Leeann Hackworth, Leisa Lutz, Ricardo Hackworth, Alvaro Smock, Dominique Smock. ## Attributes The date of birth of Gene Smock is 0898-08-16. The occupation of Gene Smock is immunologist. The hobby of Gene Smock is architecture.
Thought 4: The brother of Gene Smock is Eli Smock, and Dominique Smock has no brother. Since Gene Smock and Dominique Smock are the parents of William Smock, the uncle of William Smock is Eli Smock. So the answer is Eli Smock.
Action 4: Finish[Eli Smock].

Example 6:
Question: What is the occupation of the sister of the grandmother of Virgil Hackworth?
Thought 1: First, I need to retrieve article about Virgil Hackworth. To find the grandmother of Virgil Hackworth, I need to find the mother of his parent.
Action 1: RetrieveArticle[Virgil Hackworth].
Observation 1: # Virgil Hackworth ## Family The parents of Virgil Hackworth are Ricardo Hackworth, Vicki Hackworth. Virgil Hackworth's siblings are Leeann

Hackworth, Leisa Lutz. The sisters of Virgil Hackworth are Leeann Hackworth, Leisa Lutz. The mother of Virgil Hackworth is Vicki Hackworth. The father of Virgil Hackworth is Ricardo Hackworth. ## Friends The friends of Virgil Hackworth are Alison Smock, Eli Smock, Orlando Beltran, Ryan Wang. ## Attributes The date of birth of Virgil Hackworth is 1009-06-23. The occupation of Virgil Hackworth is sound technician. The hobby of Virgil Hackworth is wikipedia editing.
Thought 2: The parents of Virgil Hackworth are Ricardo Hackworth, Vicki Hackworth. To find the grandmother of Virgil Hackworth, I need to find the mother of Ricardo Hackworth and Vicki Hackworth.
Action 2: RetrieveArticle[Ricardo Hackworth].
Observation 2: # Ricardo Hackworth ## Family The children of Ricardo Hackworth are Leeann Hackworth, Leisa Lutz, Virgil Hackworth. The son of Ricardo Hackworth is Virgil Hackworth. The daughters of Ricardo Hackworth are Leeann Hackworth, Leisa Lutz. The wife of Ricardo Hackworth is Vicki Hackworth. ## Friends The friends of Ricardo Hackworth are Gene Smock, Isiah Lutz, Johnetta Wang. ## Attributes The date of birth of Ricardo Hackworth is 0983-02-24. The occupation of Ricardo Hackworth is clinical research associate. The hobby of Ricardo Hackworth is dairy farming.
Thought 3: Ricardo Hackworth has no mother. So I need to retrieve article about Vicki Hackworth and find her mother.
Action 3: RetrieveArticle[Vicki Hackworth].
Observation 3: # Vicki Hackworth ## Family The parents of Vicki Hackworth are Dino Beltran, Shelli Beltran. Vicki Hackworth's siblings are Aida Wang, Barabara Beltran. The sisters of Vicki Hackworth are Aida Wang, Barabara Beltran. The mother of Vicki Hackworth is Shelli Beltran. The father of Vicki Hackworth is Dino Beltran. The children of Vicki Hackworth are Leeann Hackworth, Leisa Lutz, Virgil Hackworth. The son of Vicki Hackworth is Virgil Hackworth. The daughters of Vicki Hackworth are Leeann Hackworth, Leisa Lutz. The husband of Vicki Hackworth is Ricardo Hackworth. ## Friends The friends of Vicki Hackworth are Brian Beltran, Dominique Smock, Eli Smock. ## Attributes The date of birth of Vicki Hackworth is 0985-05-30. The occupation of Vicki Hackworth is police officer. The hobby of Vicki Hackworth is meditation.
Thought 4: The mother of Vicki Hackworth is Shelli Beltran. So the only grandmother of Virgil Hackworth is Shelli Beltran. Now I need to retrieve article about Shelli Beltran and find her sister.
Action 4: RetrieveArticle[Shelli Beltran].
Observation 4: # Shelli Beltran ## Family The parents of Shelli Beltran are Alison Smock, Williams Smock. Shelli Beltran's sibling is Stacia Toombs. The sister of Shelli Beltran is Stacia Toombs. The mother of Shelli Beltran is Alison Smock. The father of Shelli Beltran is Williams Smock. The children of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth. The daughters of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth. The husband of Shelli Beltran is Dino Beltran. ## Friends The friends of Shelli Beltran are Brian Beltran, Eli Smock, Isiah Lutz, Leslee Toombs, Lesley Lutz, Ryan Wang. ## Attributes The date of birth of Shelli Beltran is 0958-03-08. The occupation of Shelli Beltran is occupational therapist. The hobby of Shelli Beltran is sociology.
Thought 5: The sister of Shelli Beltran is Stacia Toombs. Now I need to retrieve article about Stacia Toombs and find her occupation.
Action 5: RetrieveArticle[Stacia Toombs].
Observation 5: # Stacia Toombs ## Family The parents of Stacia Toombs are Alison Smock, Williams Smock. Stacia Toombs's sibling is Shelli Beltran. The

```
   sister of Stacia Toombs is Shelli Beltran.  The mother of Stacia Toombs is
   Alison Smock.  The father of Stacia Toombs is Williams Smock.  The child of
   Stacia Toombs is Leslee Toombs.  The daughter of Stacia Toombs is Leslee
   Toombs.  The husband of Stacia Toombs is Wilbert Toombs.  ## Friends The
   friends of Stacia Toombs are Brian Beltran, Isiah Lutz, Leeann Hackworth,
   Lesley Lutz, Ryan Wang.  ## Attributes The date of birth of Stacia Toombs is
   0959-03-22.  The occupation of Stacia Toombs is actuary.  The hobby of Stacia
    Toombs is finance.
Thought 6: The occupation of Stacia Toombs is actuary. So the answer is actuary.
Action 6: Finish[actuary].

Example 7:
Question: Who is the brother of the person whose occupation is associate
   professor?
Thought 1: First, I need to search for people whose occupation is associate
   professor.
Action 1: Search[associate professor].
Observation 1: (1) Dino Beltran
Thought 2: Person whose hobby is associate professor is Dino Beltran. Now I need
   to retrieve article about Dino Beltran and find his brother.
Action 2: RetrieveArticle[Dino Beltran].
Observation 2: # Dino Beltran ## Family The parents of Dino Beltran are Brian
   Beltran, Daisy Beltran.  Dino Beltran's sibling is Orlando Beltran.  The
   brother of Dino Beltran is Orlando Beltran.  The mother of Dino Beltran is
   Daisy Beltran.  The father of Dino Beltran is Brian Beltran.  The children of
    Dino Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth.  The
   daughters of Dino Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth.
   The wife of Dino Beltran is Shelli Beltran.  ## Friends The friend of Dino
   Beltran is Alvaro Smock.  ## Attributes The date of birth of Dino Beltran is
   0958-08-09.  The occupation of Dino Beltran is associate professor.  The
   hobby of Dino Beltran is shogi.
Thought 3: The brother of Dino Beltran is Orlando Beltran, so the answer is
   Orlando Beltran.
Action 3: Finish[Orlando Beltran].

Example 8:
Question: What is the date of birth of the person whose hobby is meteorology?
Thought 1: First, I need to search for people whose hobby is meteorology.
Action 1: Search[meteorology].
Observation 1: (1) Alison Smock (2) Barabara Beltran
Thought 2: People whose hobby is meteorology are Alison Smock, Barabara Beltran.
   Now I need to retrieve article about Alison Smock and Barabara Beltran and
   find their date of birth.
Action 2: RetrieveArticle[Alison Smock].
Observation 2: # Alison Smock ## Family The children of Alison Smock are Shelli
   Beltran, Stacia Toombs.  The daughters of Alison Smock are Shelli Beltran,
   Stacia Toombs.  The husband of Alison Smock is Williams Smock.  ## Friends
   The friends of Alison Smock are Eli Smock, Lannie Smock, Virgil Hackworth.
   ## Attributes The date of birth of Alison Smock is 0929-10-28.  The
   occupation of Alison Smock is health promotion specialist.  The hobby of
   Alison Smock is meteorology.
Thought 3: The date of birth of Alison Smock is 0929-10-28. Now I need to
   retrieve article about Barabara Beltran and find her date of birth.
Action 3: RetrieveArticle[Barabara Beltran].
```

Observation 3: # Barabara Beltran ## Family The parents of Barabara Beltran are Dino Beltran, Shelli Beltran. Barabara Beltran's siblings are Aida Wang, Vicki Hackworth. The sisters of Barabara Beltran are Aida Wang, Vicki Hackworth. The mother of Barabara Beltran is Shelli Beltran. The father of Barabara Beltran is Dino Beltran. ## Friends ## Attributes The date of birth of Barabara Beltran is 0989-06-11. The occupation of Barabara Beltran is broadcast engineer. The hobby of Barabara Beltran is meteorology.
Thought 4: The date of birth of Barabara Beltran is 0989-06-11. So the answer is 0929-10-28, 0989-06-11.
Action 4: Finish[0929-10-28{constants.answer_sep}0989-06-11].

Example 9:
Question: Who is the cousin of the person whose occupation is broadcast engineer?
Thought 1: First, I need to search for people whose occupation is broadcast engineer.
Action 1: Search[broadcast engineer].
Observation 1: # Barabara Beltran ## Family The parents of Barabara Beltran are Dino Beltran, Shelli Beltran. Barabara Beltran's siblings are Aida Wang, Vicki Hackworth. The sisters of Barabara Beltran are Aida Wang, Vicki Hackworth. The mother of Barabara Beltran is Shelli Beltran. The father of Barabara Beltran is Dino Beltran. ## Friends ## Attributes The date of birth of Barabara Beltran is 0989-06-11. The occupation of Barabara Beltran is broadcast engineer. The hobby of Barabara Beltran is meteorology.
Thought 2: Person whose hobby is broadcast engineer is Barabara Beltran. The cousin of Barabara Beltran is the child of sibling of her parent. Barabara Beltran's parents are Dino Beltran and Shelly Beltran. Now I need to find the sibling of Dino Beltran and Shelly Beltran.
Action 2: RetrieveArticle[Dino Beltran].
Observation 2: # Dino Beltran ## Family The parents of Dino Beltran are Brian Beltran, Daisy Beltran. Dino Beltran's sibling is Orlando Beltran. The brother of Dino Beltran is Orlando Beltran. The mother of Dino Beltran is Daisy Beltran. The father of Dino Beltran is Brian Beltran. The children of Dino Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth. The daughters of Dino Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth. The wife of Dino Beltran is Shelli Beltran. ## Friends The friend of Dino Beltran is Alvaro Smock. ## Attributes The date of birth of Dino Beltran is 0958-08-09. The occupation of Dino Beltran is associate professor. The hobby of Dino Beltran is shogi.
Thought 3: The sibling of Dino Beltran is Orlando Beltran. Now I need to retrieve article about Orlando Beltran and find his child.
Action 3: RetrieveArticle[Orlando Beltran].
Observation 3: # Orlando Beltran ## Family The parents of Orlando Beltran are Brian Beltran, Daisy Beltran. Orlando Beltran's sibling is Dino Beltran. The brother of Orlando Beltran is Dino Beltran. The mother of Orlando Beltran is Daisy Beltran. The father of Orlando Beltran is Brian Beltran. ## Friends The friends of Orlando Beltran are Virgil Hackworth, Dominique Smock. ## Attributes The date of birth of Orlando Beltran is 0953-10-23. The occupation of Orlando Beltran is petroleum engineer. The hobby of Orlando Beltran is learning.
Thought 4: Orlando Beltran has no child, so no cousin from the side of Dino Beltran. Now I need to retrieve article about Shelli Beltran and find her sibling.
Action 4: RetrieveArticle[Shelli Beltran].

```
Observation 4: # Shelli Beltran ## Family The parents of Shelli Beltran are
   Alison Smock, Williams Smock.  Shelli Beltran's sibling is Stacia Toombs.
   The sister of Shelli Beltran is Stacia Toombs.  The mother of Shelli Beltran
   is Alison Smock.  The father of Shelli Beltran is Williams Smock.  The
   children of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki Hackworth.
    The daughters of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki
   Hackworth.  The husband of Shelli Beltran is Dino Beltran.  ## Friends The
   friends of Shelli Beltran are Brian Beltran, Eli Smock, Isiah Lutz, Leslee
   Toombs, Lesley Lutz, Ryan Wang.  ## Attributes The date of birth of Shelli
   Beltran is 0958-03-08.  The occupation of Shelli Beltran is occupational
   therapist.  The hobby of Shelli Beltran is sociology.
Thought 5: The sibling of Shelli Beltran is Stacia Toombs. Now I need to retrieve
    article about Stacia Toombs and find her child, to get the cousin of
   Barabara Beltran.
Action 5: RetrieveArticle[Stacia Toombs].
Observation 5: # Stacia Toombs ## Family The parents of Stacia Toombs are Alison
   Smock, Williams Smock.  Stacia Toombs's sibling is Shelli Beltran.  The
   sister of Stacia Toombs is Shelli Beltran.  The mother of Stacia Toombs is
   Alison Smock.  The father of Stacia Toombs is Williams Smock.  The child of
   Stacia Toombs is Leslee Toombs.  The daughter of Stacia Toombs is Leslee
   Toombs.  The husband of Stacia Toombs is Wilbert Toombs.  ## Friends The
   friends of Stacia Toombs are Brian Beltran, Isiah Lutz, Leeann Hackworth,
   Lesley Lutz, Ryan Wang.  ## Attributes The date of birth of Stacia Toombs is
   0959-03-22.  The occupation of Stacia Toombs is actuary.  The hobby of Stacia
    Toombs is finance.
Thought 6: The child of Stacia Toombs is Leslee Toombs. So the cousin of Barabara
    Beltran is Leslee Toombs. The answer is Leslee Toombs.
Action 6: Finish[Leslee Toombs].

Example 10:
Question: Who is the great-granddaughter of the person whose hobby is biology?
Thought 1: First, I need to search for people whose hobby is biology.
Action 1: Search[biology].
Observation 1: (1) Alvaro Smock
Thought 2: Person whose hobby is biology is Alvaro Smock. Now I need to retrieve
   article about Alvaro Smock and find his great-granddaughter. A great-
   granddaughter is the daughter of the child of the child of the person.
Action 2: RetrieveArticle[Alvaro Smock].
Observation 2: # Alvaro Smock ## Family The children of Alvaro Smock are Eli
   Smock, Gene Smock.  The sons of Alvaro Smock are Eli Smock, Gene Smock.  The
   wife of Alvaro Smock is Lannie Smock.  ## Friends The friends of Alvaro Smock
    are Dino Beltran, Gene Smock, Aida Wang.  ## Attributes The date of birth of
    Alvaro Smock is 0867-07-12.  The occupation of Alvaro Smock is osteopath.
   The hobby of Alvaro Smock is biology.
Thought 3: The children of Alvaro Smock are Eli Smock, Gene Smock. Now I need to
   retrieve article about Eli Smock and Gene Smock and find their child.
Action 3: RetrieveArticle[Eli Smock].
Observation 3: # Eli Smock ## Family The parents of Eli Smock are Alvaro Smock,
   Lannie Smock.  Eli Smock's sibling is Gene Smock.  The brother of Eli Smock
   is Gene Smock.  The mother of Eli Smock is Lannie Smock.  The father of Eli
   Smock is Alvaro Smock.  ## Friends The friends of Eli Smock are Leisa Lutz,
   Shelli Beltran, Vicki Hackworth, Virgil Hackworth, Alison Smock, Brian
   Beltran.  ## Attributes The date of birth of Eli Smock is 0901-01-18.  The
   occupation of Eli Smock is retail manager.  The hobby of Eli Smock is tether
```

Table 2. **F1 scores (in %) for Llama-3.3-70B with RAG prompting.** ZEROSHOT-RAG and COT-RAG use the UAE-LARGE-V1 retriever model (top-k= 4); SELF-ASK and IRCOT use the BM25 retriever (top-k= 5). We report mean ± 1 standard error across 3 dataset generation seeds.

Universe Size n	ZEROSHOT-RAG	COT-RAG	SELF-ASK	IRCOT
50	17.55 ± 2.20	20.01 ± 1.81	13.41 ± 0.98	23.93 ± 0.93

```
    car.
Thought 4: Eli Smock has no child. Now I need to retrieve article about Gene
    Smock and find his child.
Action 4: RetrieveArticle[Gene Smock].
Observation 4: # Gene Smock ## Family The parents of Gene Smock are Alvaro Smock,
     Lannie Smock.  Gene Smock's sibling is Eli Smock.  The brother of Gene Smock
     is Eli Smock.  The mother of Gene Smock is Lannie Smock.  The father of Gene
     Smock is Alvaro Smock.  The child of Gene Smock is Williams Smock.  The son
    of Gene Smock is Williams Smock.  The wife of Gene Smock is Dominique Smock.
     ## Friends The friends of Gene Smock are Leeann Hackworth, Leisa Lutz,
    Ricardo Hackworth, Alvaro Smock, Dominique Smock.  ## Attributes The date of
    birth of Gene Smock is 0898-08-16.  The occupation of Gene Smock is
    immunologist.  The hobby of Gene Smock is architecture.
Thought 5: The child of Gene Smock is Williams Smock. Now I need to retrieve
    article about Williams Smock and find his daughter, to get the great-
    granddaughter of Alvaro Smock.
Action 5: RetrieveArticle[Williams Smock].
Observation 5: # Williams Smock ## Family The parents of Williams Smock are
    Dominique Smock, Gene Smock.  The mother of Williams Smock is Dominique Smock
    .  The father of Williams Smock is Gene Smock.  The children of Williams
    Smock are Shelli Beltran, Stacia Toombs.  The daughters of Williams Smock are
     Shelli Beltran, Stacia Toombs.  The wife of Williams Smock is Alison Smock.
     ## Friends The friend of Williams Smock is Lannie Smock.  ## Attributes The
    date of birth of Williams Smock is 0926-04-04.  The occupation of Williams
    Smock is clinical biochemist.  The hobby of Williams Smock is social studies.
Thought 6: The daughters of Williams Smock are Shelli Beltran, Stacia Toombs. So
    the great-granddaughters of Alvaro Smock, whose hobby is biology, are Shelli
    Beltran, Stacia Toombs. The answer is Shelli Beltran, Stacia Toombs.
Action 6: Finish[Shelli Beltran{constants.answer_sep}Stacia Toombs].
(END OF EXAMPLES)

Now answer the following question:
Question: {{question}}
{{scratchpad}}
```

D. Additional Experiments

D.1. Article Rephrasing

To explore how LLMs can be used to improve the realism of our templated articles, we instruct Llama-3.3-70B to rephrase the corpora of size $n = 50$ from Figure 2. Our first, "short" prompt instructs the LLM to condense the templated articles, while still retaining all factual information:

```
Shuffle and rephrase the following wikipedia-like article. Keep ALL facts
like name, relation, date, occupation, hobby and gender exactly as stated. Do
not add ANY new information.
```

Our second, "long" prompt permits the LLM to introduce new facts, without contradicting existing facts:

```
Transform the following factual article into an engaging narrative profile:

1. Add colorful descriptions and personality traits that might be inferred
2. Create a vivid backstory about how relationships formed
3. Elaborate on the person's career path and achievements
4. Describe their hobbies in rich detail, including when they might have started
   them
5. Include hypothetical quotes from family members
6. Imagine and describe the person's daily routine
7. Add details about where they might live and their home environment
8. Write in a warm, personal tone as if you've known the subject for years
9. Maintain a natural flow

When transforming the article, ensure the following:
```

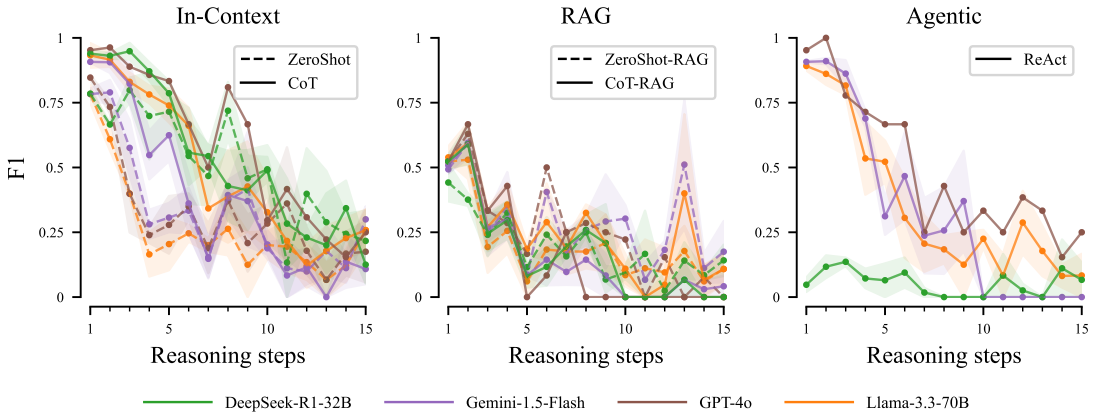


Figure 5. **F1 scores versus question difficulty, measured by *reasoning steps*, for questions with exactly 1 solution.** We observe similar trends as in Figure 2, demonstrating that the number of solutions is not solely responsible for the drop in performance.

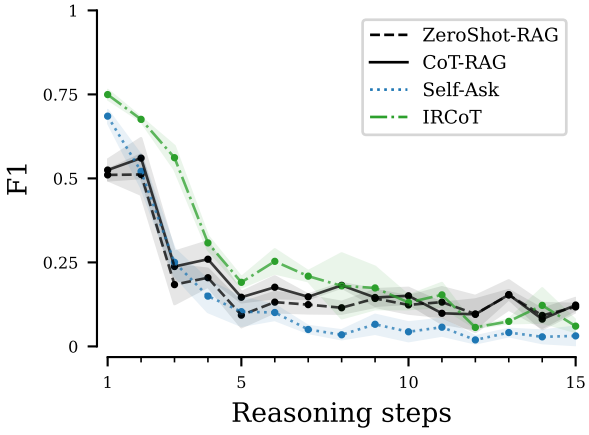
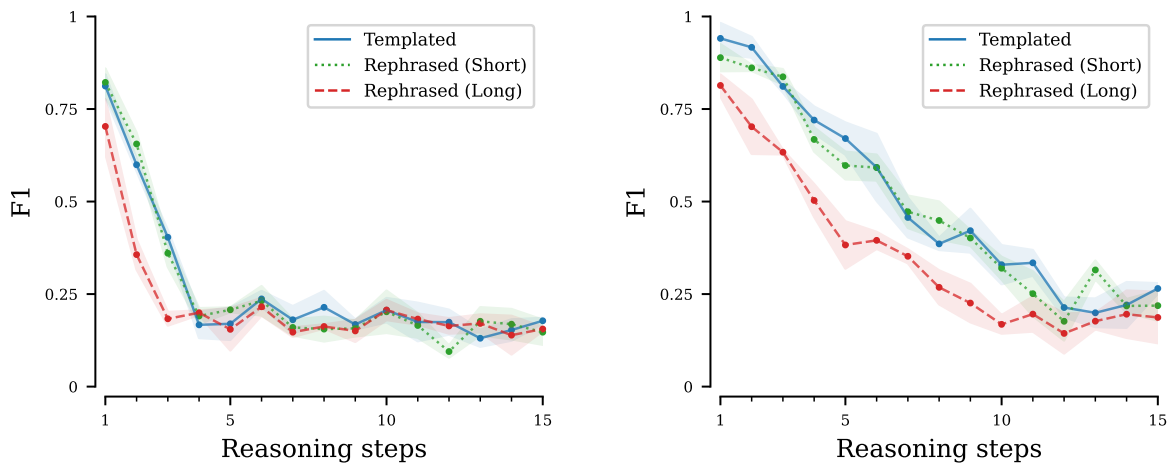


Figure 6. **F1 scores versus question difficulty, measured by reasoning steps.** ZEROSHOT-RAG and COT-RAG use the UAE-LARGE-V1 retriever model (top-k= 4); SELF-ASK and IRCOT use the BM25 retriever (top-k= 5). For all methods, we use Llama-3.3-70B as the generator model. We report mean $\pm$ 1 standard error across 3 dataset generations seeds with universe size $n = 50$.



(a) ZEROSHOT Prompting (b) COT Prompting

Figure 7. **F1 scores versus difficulty, measured by reasoning steps, for various article styles**. For rephrasing, we use Llama-3.3-70B with temperature= 0.7 and top-p= 0.9. For evaluation, we use Llama-3.3-70B with the ZEROSHOT and COT prompts from Appendix C.2 and Appendix C.4, respectively, and the sampling hyper-parameters from Appendix C.1. As in Figure 2, we report mean ± 1 standard error across 3 dataset generation seeds for universe size $n = 50$.

```
1. All factual information remains 100% accurate (names, relationships, dates,
   occupation, hobby, gender)
2. No new information is added or implied
3. The content is presented in a natural, flowing narrative rather than just
   shuffling bullet points
4. Different sections are reorganized in a way that still makes logical sense
5. Sentence structures and vocabulary are varied from the original

This is the article to transform:
```

For both prompts, we use temperature = 0.7, top-p = 0.9, and max output tokens = 1024. We include an example short article in Appendix D.1.1 and an example long article in Appendix D.1.3. Figure 7 shows the F1 scores of Llama-3.3-70B using ZEROSHOT and COT prompting on these rephrased articles. Remarkably, we observe that F1 declines when the number of the reasoning steps increases, regardless of the style of articles.

D.1.1. EXAMPLE OF TEMPLATED ARTICLE

```
# 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 Cortney 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.
```

D.1.2. Example of Short Rephrased Article

The following article was generated with Llama-3.3-70B using 118 output tokens.

> Alison Smock, a female born on 0929-10-28, is a broadcast engineer with a hobby of meteorology. She is married to Williams Smock and they have daughters named Cythia Smock, Shelli Beltran, and Stacia Toombs. Alison Smock's social circle includes friends such as Cortney Parmer, Jamison Baptiste, and Wilbert Toombs.

D.1.3. Example of Long Rephrased Article

The following article was generated with Llama-3.3-70B using 669 output tokens. Llama-3.3-70B incorrectly stated the date-of-birth (red) and omitted the last names of the friends (orange).

> 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. 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, Alison loved to indulge in her hobby of meteorology. She found the study of weather patterns and atmospheric conditions absolutely captivating, and she spent hours poring over books and watching documentaries on the subject. As her friend Cortney Parmer recalls, "Alison would often call me up and say, 'Did you see that storm system moving in? I've been tracking it all day!' She was like a kid in a candy store when it came to weather."
>
> In her free time, Alison enjoyed spending time with her friends, including Cortney, Jamison Baptiste, and Wilbert Toombs. They would often meet up for coffee or go on hikes together, discussing everything from the latest news to their personal lives. Wilbert, who is also the husband of Alison's daughter Stacia, says, "Alison is like a mother to me. She's always been so kind and supportive, and I admire her intelligence and passion for life."
>
> A typical day for Alison begins early, with a quiet moment of reflection and a cup of coffee on her back porch, surrounded by the beauty of nature. She lives in a cozy home filled with books, gadgets, and mementos from her travels, and she loves to spend time in her garden, tending to her plants and watching the birds. As she gets ready for work, Alison feels grateful for the life she's built and the people who bring joy and love into her world. As her husband Williams says, "Alison is an amazing woman, and I feel so lucky to have her by my side. She's intelligent, kind, and beautiful, and I admire her every day." With her quick wit, sharp mind, and generous spirit, Alison Smock is a true treasure, and those who know her feel fortunate to have her in their lives.

D.2. Fine-Tuning on PhantomWiki

While PhantomWiki is meant for evaluation, we also assess the viability of PhantomWiki for training language models. We generate 10 new PhantomWiki dataset instances (question depth 20 and universe size 50) amounting to 5K training question-answer pairs. We then perform full fine-tuning of Qwen2.5-0.5B-Instruct and parameter-efficient fine-tuning of Qwen2.5-3B-Instruct with LoRA (Hu et al., 2022) applied to all linear layers.

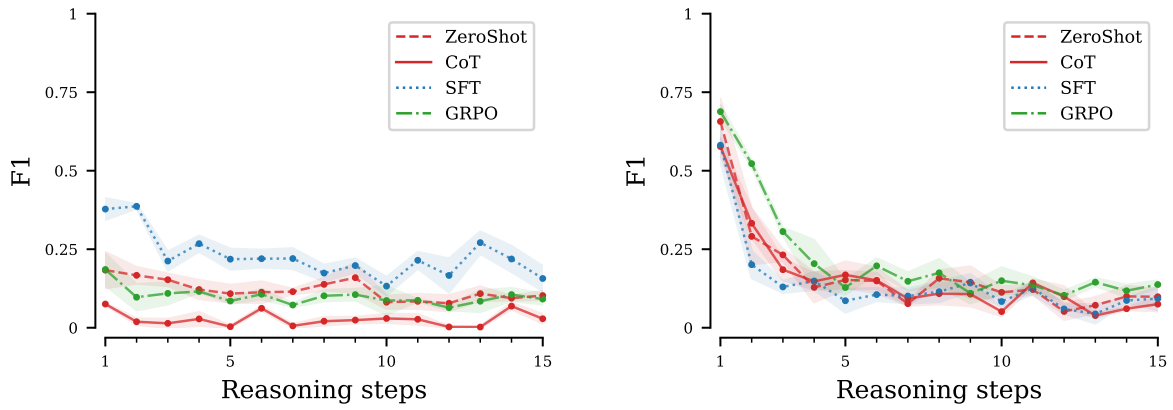
For each base model, we employ two popular training algorithms. The first is Group Relative Policy Optimization (GRPO) from Shao et al. (2024b) using an F1-score reward. We use the CoT prompt template in Appendix C.4 with batch size 32. The second is supervised fine-tuning (SFT) with the answer as the ground-truth label. We use the ZeroShot prompt in

Table 3. **F1 scores (in %).** We report the mean ± 1 standard error across 3 dataset generation seeds for universe size $n = 50$. ZEROSHOT and COT are prompting methods; SFT and GRPO are fine-tuning methods.

Method	Qwen2.5-0.5B	Qwen2.5-3B
ZEROSHOT	11.84 ± 0.82	16.65 ± 2.37
COT	2.55 ± 0.28	14.84 ± 1.12
SFT	22.72 ± 0.48	13.30 ± 1.58
GRPO	9.70 ± 0.45	20.46 ± 0.45

Appendix C.2 and batch size 4. For all training experiments, we train for 3 epochs until convergence using the AdamW optimizer with initial learning rate set to 5×10^{-6} for full fine-tuning and 10^{-4} for LoRA fine-tuning.

We evaluate these models on the three PhantomWiki instances of size $n = 50$ and maximum recursion depth 20 used in Figure 2. Figure 8 compares the F1 score of prompting versus fine-tuning across difficulty levels, measured by reasoning steps, while Table 3 contains the mean F1 scores. For Qwen2.5-0.5B-Instruct, we see that GRPO and SFT improves performance compared to COT and ZEROSHOT prompting, respectively. For Qwen2.5-3B-Instruct, we see that GRPO improves performance over to the COT baseline, but SFT worsens performance over to the ZEROSHOT baseline.



(a) Qwen2.5-0.5B-Instruct (full fine-tuning)

(b) Qwen2.5-3B-Instruct (LoRA fine-tuning)

Figure 8. **F1 scores versus difficulty, measured by reasoning steps.** We report mean ± 1 standard error across 3 dataset generation seeds for universe size $n = 50$. ZEROSHOT and COT are prompting methods; SFT and GRPO are fine-tuning methods.

E. Example of Small Corpus

We generate a universe of size $n = 4$, setting the number of family trees to be one. We include the articles below.

E.1. Article 1

```
# Claud Colin

## Family
The brother of Claud Colin is Mckinley Colin.
The mother of Claud Colin is Ramona Colin.
The father of Claud Colin is Danilo Colin.

## Friends
The friend of Claud Colin is Danilo Colin.

## Attributes
```

```
The date of birth of Claud Colin is 0241-12-06.
The occupation of Claud Colin is academic librarian.
The hobby of Claud Colin is amateur astronomy.
The gender of Claud Colin is male.
```

E.2. Article 2

```
# Danilo Colin

## Family
The sons of Danilo Colin are Claud Colin, Mckinley Colin.
The wife of Danilo Colin is Ramona Colin.

## Friends
The friends of Danilo Colin are Mckinley Colin, Ramona Colin, Claud Colin.

## Attributes
The date of birth of Danilo Colin is 0219-08-09.
The occupation of Danilo Colin is clinical research associate.
The hobby of Danilo Colin is crystals.
The gender of Danilo Colin is male.
```

E.3. Article 3

```
# Mckinley Colin

## Family
The brother of Mckinley Colin is Claud Colin.
The mother of Mckinley Colin is Ramona Colin.
The father of Mckinley Colin is Danilo Colin.

## Friends
The friends of Mckinley Colin are Ramona Colin, Danilo Colin.

## Attributes
The date of birth of Mckinley Colin is 0246-10-18.
The occupation of Mckinley Colin is museum curator.
The hobby of Mckinley Colin is stamp collecting.
The gender of Mckinley Colin is male.
```

E.4. Article 4

```
# Ramona Colin

## Family
The sons of Ramona Colin are Claud Colin, Mckinley Colin.
The husband of Ramona Colin is Danilo Colin.

## Friends
The friends of Ramona Colin are Danilo Colin, Mckinley Colin.

## Attributes
The date of birth of Ramona Colin is 0219-09-08.
The occupation of Ramona Colin is technical sales engineer.
The hobby of Ramona Colin is trainspotting.
The gender of Ramona Colin is female.
```