Identifying Where Large Language Models Struggle in Answering Complex Questions

Xanh Ho,[1] **Florian Boudin,**[2] **Saku Sugawara,**[1] **Khoa Duong,**[3] and **Akiko Aizawa**[1]
[1]National Institute of Informatics, Japan
[2]Inria, LS2N, Nantes Université, France [3]Independent Researcher
{xanh, saku, aizawa}@nii.ac.jp
florian.boudin@univ-nantes.fr dnanhkhoa@live.com

Abstract

We design experiments to identify where Large Language Models (LLMs) struggle when answering complex questions. Our focus is on two key stages, mirroring the human QA process: 1) *question decomposition*, where the model breaks down a complex question into sub-questions and 2) *subproblem solving*, where it addresses each sub-question to obtain the final response. We preprocess and expand three multi-hop datasets to create experimental datasets featuring explicit and implicit multi-hop questions, crowdsourced and templated questions, and varying numbers of hops. Our results show that larger models (Llama 3.1 70B and o1) excel at decomposing explicit multi-hop questions but struggle with implicit ones, while smaller models (e.g., Llama 3.1 8B) have difficulty with both. In the sub-problem solving stage, all models perform well on simple questions with context. Furthermore, we found no correlation between accuracy in the question decomposition stage and final QA performance (direct response), highlighting a key difference between human and LLM reasoning.[1]

1 Introduction

With the release of language models (LMs) (Devlin et al., 2019; Yang et al., 2019, *inter alia*), especially large language models (LLMs) (Brown et al., 2020; Zhao et al., 2023, *inter alia*), there has been a significant change in the research community. Using well-designed prompts (e.g., chain-of-thought (CoT; Wei et al., 2022)), previous studies (Dua et al., 2022; Kojima et al., 2022) have demonstrated the reasoning abilities of LLMs in performing various tasks, such as arithmetic reasoning (Cobbe et al., 2021), multi-hop (compositional) reasoning (Welbl et al., 2018), and commonsense reasoning (Talmor et al., 2019). While LLMs have demonstrated impressive performance on many complex question-answering (QA) tasks, our understanding of their answering process remains limited.

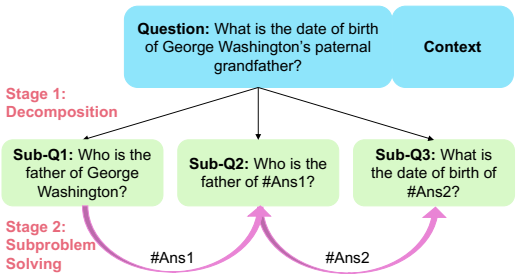

Figure 1: We examine model behavior across two stages: 1) question decomposition and 2) subproblem solving.

From a human perspective, complex questions are typically answered through a process of decomposition (Pelletier, 1994; Wolfson et al., 2020). For example, to answer the question [*What is the date of birth of George Washington's paternal grandfather?*] (as shown in Figure 1), humans would break it down into a chain of simpler sub-questions, beginning with [*Who is the father of George Washington?*]. They would then answer these sub-questions to obtain the final answer (Talmor and Berant, 2018; Perez et al., 2020; Khot et al., 2023). Previous research has employed question decomposition in the QA process to enhance model performance and explainability (Min et al., 2019b; Fu et al., 2021; Press et al., 2023; Zhou et al., 2023; Zhang et al., 2024). However, little prior work appears to explore the ability of LLMs to replicate the two-stage process humans use to answer complex questions, particularly with annotations for validating decomposed questions.

In this work, we aim to explore the ability of LLMs to answer complex questions by mimicking the human QA process. Specifically, we examine how models perform across two key stages: 1) *question decomposition*, where the model breaks down a complex question into a chain of sub-questions and 2) *subproblem solving*, where it addresses each

[1]Our data and code are available at https://github.com/Alab-NII/complex_ques_decomposition

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 112–123

sub-question to obtain the final response. An example of these stages is illustrated in Figure 1. Since LLMs operate as black-boxes, we argue that identifying where they fail throughout the process of answering complex questions will provide valuable insights for improving their effectiveness.

To gain a broader perspective on LLM capabilities, we experiment with three distinct multi-hop reasoning datasets: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (2Wiki; Ho et al., 2020), and StrategyQA (Geva et al., 2021). These datasets feature both explicit and implicit multi-hop questions, sourced from human crowdsourcing and template-based generation, with the number of hops varying from 2 to 4. We conduct our investigations using three LLMs of different sizes –small, medium and large– from two families: Llama 3.1 and GPT (o1).

Our experimental results indicate that while larger models (e.g., Llama 3.1 70B and o1) effectively decompose explicit multi-hop questions, they struggle with implicit ones, whereas smaller models (such as Llama 3.1 8B) face challenges with both types. In the sub-problem solving stage, all models demonstrate strong performance when answering simple questions given the context. Additionally, we observed no clear correlation between question decomposition accuracy and final QA performance, underscoring a fundamental difference between human and LLM reasoning. Humans often explicitly decompose complex questions into smaller, manageable steps, whereas LLMs may instead rely on latent multi-hop reasoning or heuristic shortcuts, as suggested in prior work, or potentially other mechanisms that have not yet been discovered by the research community.

2 Related Work

The reasoning process with decomposition steps in multi-hop QA offers benefits such as supporting explanation-focused evaluation tasks (Ho et al., 2020; Inoue et al., 2020; Tang et al., 2021) or improving answer explainability and final QA performance when integrated into model design (Min et al., 2019b; Fu et al., 2021; Press et al., 2023; Radhakrishnan et al., 2023; Zhou et al., 2023). Some previous works (Wu et al., 2024; Han and Gardent, 2025; Ammann et al., 2025) attempt to propose a model or system to decompose questions into subquestions. Recently, several studies have analyzed the internal workings of LLMs and shown that latent multi-hop reasoning may exist (Yang et al., 2024; Biran et al., 2024; Yang et al., 2025). Our work differs from prior studies by examining whether LLMs replicate the human two-stage process for answering complex questions.

3 Datasets

We focus on the multi-hop reasoning task (Welbl et al., 2018), which requires models to gather information from multiple paragraphs, and then perform compositional reasoning to obtain the final answer. To broaden our understanding of LLM capabilities, we select multi-hop reasoning datasets with explicit and implicit questions, sourced from crowdsourcing and templates, and hop counts ranging from 2 to 4. We decided to choose HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), and StrategyQA (Geva et al., 2021). Appendix A shows the per-hop and total sample counts for each dataset.

HotpotQA. HotpotQA questions are crowdsourced based on two gold paragraphs, with most being 2-hop. As the original HotpotQA lacks sub-questions and sub-answers, we use a 1,000-question subset from Tang et al. (2021) that includes them. To filter samples likely to involve reasoning shortcuts (e.g., overlap between the first sub-answer and the final answer), we apply automated criteria, yielding 874 samples. We reserve 87 (10%) for development (e.g., parameter selection) and use the remaining 787 for evaluation.

2Wiki. 2Wiki questions are template-based and mostly two-hop. Since template-based questions are easy to expand, we develop a process to extend them to 3-hop and 4-hop while ensuring answerability. The extension process involves four steps: (1) obtaining basic 2-hop samples, (2) manually verifying samples and related Wikidata triples, (3) preparing templates, and (4) generating samples automatically. Details of our process are provided in Appendix A.1. Following standard practice, we reserve 10% (84 samples) for development and 756 for evaluation. We refer to the extended version used in our experiment as 2Wiki-complex.

StrategyQA. StrategyQA questions are crowdsourced. Unlike HotpotQA and 2Wiki, where the reasoning process is explicitly stated in the question, StrategyQA questions are implicit, with all reasoning steps unstated (Geva et al., 2021). We select 620 training samples with 2–4 sub-questions and 2–4 unique paragraphs. Since sub-answers

Dataset	Llama 3.1 8B					Llama 3.1 70B					o1				
	Δsub	linked	PD	RC	CSQ	Δsub	linked	PD	RC	CSQ	Δsub	linked	PD	RC	CSQ
Hotpot	0.3	64.9	3.2	9.5	31.3	71.1	97.3	57.9	64.2	67.5	80.9	97.9	63.2	68.4	70.7
2Wiki-c	3.2	79.8	4.0	5.0	33.2	81.0	100.0	85.0	85.0	93.8	92.0	100.0	87.0	91.0	92.3
Strategy	14.8	17.8	5.4	8.7	30.3	39.1	79.7	41.3	45.7	63.2	58.4	97.8	19.6	20.7	32.2

Table 1: Automatic and human scores (green) in the decomposition stage. PD, RC, and CSQ denote *Perfect Decomposition*, *Reasoning Chain*, and *Correct Sub-questions*, as defined in our human evaluation guidelines.

are missing, we generate them using Llama-3.3-70B based on the provided short facts. Since sub-questions are dependent, we use generated sub-answers to replace placeholders and obtain the final answer. We retain only samples where all sub-answers lead to the correct final answer and filter out those with sub-answers longer than 100 characters based on our observations. This yields 373 samples, with 37 (10%) for development and 336 for evaluation. Details are in Appendix A.2.

4 Experiments

4.1 Experimental Settings

We conduct experiments using both open-source and closed-source LLMs. For open-source, we use the instruction-tuned versions of Llama 3.1 8B, Llama 3.1 70B, and Llama 3.3 70B (Grattafiori et al., 2024). For closed-source, we use o1 (o1-2024-12-17). We run experiments on development sets and observe that 2-shot prompting often yields better automatic scores. Therefore, we use 2-shot prompting in our decomposition stage. Note that we use Llama 3.3 70B solely for judgment.

4.2 Results

4.2.1 Decomposition Stage

At this stage, the input is a complex question, and the output is a list of connected sub-questions. Since there are many valid ways to split a complex question into sub-questions, we also provide context in the prompt and instruct the models to refer to it if decomposition is challenging. We use both automatic and human evaluation for this stage, and the scores are presented in Table 1.

Automatic Evaluation. We use two metrics: **Δsub**, which scores 1 if the number of generated sub-questions matches the gold sub-questions and 0 otherwise, and **linked**, which scores 1 if all sub-questions are connected through placeholders (e.g., #1 or #Ans1), and 0 otherwise. As shown in Table 1, large models achieve high **linked** scores, demonstrating their ability to follow instructions for linking sub-questions. Large models also achieve high **Δsub** scores, but for StrategyQA, these scores are relatively low, highlighting the challenge of generating the exact number of sub-questions in this dataset. Notably, these metrics do not evaluate subquestion content.

Human Evaluation. We manually assess 100 randomly selected samples per dataset[2] based on three criteria: (1) *Perfect Decomposition*, scored as 1 if all sub-questions are logically connected, the final question leads to a correct answer, and there are no redundant sub-questions; (2) *Reasoning Chain*, scored as 1 if the decomposition is generally correct but contains some redundancy; and (3) *Number of Correct Sub-questions*, representing the count of correctly formulated sub-questions. The full human evaluation guidelines are in Appendix B.1.

As shown in Table 1, the small model (Llama 3.1 8B) performs poorly across all three datasets, indicating a lack of decomposition ability. In contrast, larger models (Llama 3.1 70B and o1) excel at decomposing explicit multi-hop questions (Hotpot and 2Wiki-c) but struggle with implicit multi-hop questions (Strategy), where the reasoning steps are not explicitly mentioned in the questions. Additionally, we observe that these larger models handle template-based multi-hop questions (2Wiki-c) well but perform less effectively on human crowd-sourced questions (Hotpot).

On StrategyQA, Llama 3.1 70B outperforms o1 by a large margin, mainly due to referencing issues in o1's outputs, such as using 'the individual in question' instead of naming entities. We manually analyze error cases across three datasets. In StrategyQA, o1 has referencing issues in 45 out of 75 cases, while Llama 3.1 70B has 1 out of 54. Accepting referencing issues, 12 cases in o1 become valid decompositions, but none in Llama 3.1 70B. As a result, o1's perfect decomposition score is 32.6,

[2]We removed 5 Hotpot and 8 Strategy samples during human evaluation due to gold decomposition issues.

Dataset	Llama 3.1 8B		Llama 3.1 70B		o1	
	Equal	All	Equal	All	Equal	All
Hotpot	96.2	92.8	96.7	93.7	94.7	90.5
2Wiki-c	95.5	86.5	99.1	97.2	99.4	98.0
Strategy	77.1	52.0	79.6	59.4	81.3	64.0

Table 2: LLM-as-a-Judge accuracy (based on Llama 3.3 70B) in the sub-problem-solving stage for two scenarios: equal weight (*Equal*) and all sub-questions correct (*All*).

Dataset	Llama 3.1 8B		Llama 3.1 70B		o1	
	0-CoT	De-Cor	0-CoT	De-Cor	0-CoT	De-Cor
Hotpot	90.5	0.058	91.6	0.125	91.6	0.004
2Wiki-c	82.0	0.096	89.0	-0.148	99.0	**0.260**
Strategy	63.0	0.084	78.3	**0.282**	80.4	0.105

Table 3: LLM-as-a-Judge accuracy (Llama 3.3 70B) for full-QA performance using zero-shot-CoT. We also highlight in blue (De-Cor) the correlation between full-QA performance and the question decomposition.

still lower than Llama 3.1 70B's 41.3. Detailed error analysis and examples are in Appendix B.3.

4.2.2 Subproblem Solving Stage

In this stage, the input is a simple question with context; the output is the answer to that question. We observe that both Exact Match (EM) and F1 scores fail to accurately reflect model performance due to varied answer representations (e.g., different formats for names). Examples comparing EM/F1 scores and LLM-as-a-judge are in Appendix B.2.

To address this, we design a few-shot prompting approach using Llama 3.3 70B as a judge (Verga et al., 2024) to evaluate predicted answers based on the correct answer and context. The model selects one of three labels: Correct, Incorrect, or Not sure. Table 2 presents the LLM-as-a-Judge accuracy in the sub-problem-solving stage for two scenarios: (1) assuming all sub-questions are of equal importance, and (2) requiring that all sub-questions within the same sample be answered correctly. (EM and F1 scores are presented in Appendix B.2.)

As shown in Table 2, all models score high at this stage due to the simplicity of the questions. However, Strategy scores lower than the other datasets, with a notable drop when all sub-questions in a sample must be correct. A manual analysis reveals three key reasons: (1) Strategy includes comparison questions that models often answer as 'Not sure'; (2) some sub-questions have context but no answer, leading to 'cannot answer'; and (3) the lack of gold sub-answers in Strategy means that our generated sub-answers may contain errors.

4.3 Final Performance vs. Decomposition

We first evaluate models on the multi-hop QA task using 300 samples from Section 4.2.1 with zero-shot CoT (Kojima et al., 2022). We then compute Pearson correlation coefficients between multi-hop QA performance and the decomposition stage (using Perfect Decomposition). All scores are in Table 3. The results show that all models perform well on HotpotQA and 2Wiki-c with gold context. However, performance is lower on Strategy, highlighting the challenge of its implicit multi-hop question format. We observe that, in most cases, the correlation between final QA performance and the decomposition stage is negligible (with scores below 0.19). However, Llama 3.1 70B on Strategy and o1 on 2Wiki-c show a moderate correlation, suggesting some relationship between decomposition quality and overall accuracy. Overall, these results indicate that QA performance does not strongly align with variations in decomposition quality.

We hypothesize two key factors behind these correlation scores: (1) Humans typically explicitly decompose complex questions into smaller, manageable subproblems before solving them. In contrast, LLMs often rely on implicit, end-to-end reasoning, engaging in latent multi-hop reasoning (Yang et al., 2024) or statistical correlations, rather than structured decomposition. (2) Instead of following a step-by-step reasoning process as expected, models may leverage shortcuts (Chen and Durrett, 2019; Min et al., 2019a) to arrive at answers, which affects their final QA performance.

5 Conclusion

We reuse three multi-hop datasets to create data with sub-questions, sub-answers, explicit and implicit questions, and varying hops. Using these, we analyze where LLMs fail by mimicking the human QA process of decomposition and subproblem-solving. Our results show that larger models handle explicit multi-hop questions but struggle with implicit ones, while smaller models struggle with both. All models perform well on simple questions with context. Notably, decomposition accuracy appears to have only a weak correlation with final QA performance. This suggests that, unlike humans who explicitly decompose complex questions, LLMs may instead rely on alternative reasoning strategies or other poorly understood mechanisms.

Limitations

Our research has two main limitations. The first limitation is that our study assumes access to the gold paragraph (the one containing the necessary information to find the answer). However, this does not reflect real-world scenarios where the gold paragraph is unavailable, requiring models to retrieve relevant information instead.

The second concern is our reliance on LLM-as-a-judge (few-shot prompting with Llama 3.3 70B), which may not always provide accurate judgments. Since Llama 3.3 70B is not the most advanced model, it may occasionally make errors, potentially affecting evaluation reliability.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K03231 and JST FOREST Grant Number JPMJFR232R.

Ethical Statement

HotpotQA, 2WikiMultihopQA, and StrategyQA are released under the CC BY-SA 4.0, Apache License 2.0, and MIT License, respectively, all of which permit redistribution and modification. We preprocess and extend these datasets using their publicly available versions. Human evaluation was conducted by the authors of this paper. A detailed annotation guideline was developed and followed to ensure consistency and reliability in the annotation process. The dataset does not contain any personal or sensitive information.

References

Paul J. L. Ammann, Jonas Golde, and Alan Akbik. 2025. Question decomposition for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 497–507, Vienna, Austria. Association for Computational Linguistics.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, Miami, Florida, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv:2110.14168*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing complex questions makes multi-hop QA easier and more interpretable. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 169–180, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The llama 3 herd of models. *arXiv:2407.21783*.

Kelvin Han and Claire Gardent. 2025. Generating complex question decompositions in the face of distribution shifts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1189–1211, Albuquerque, New Mexico. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

Francis Jeffry Pelletier. 1994. The principle of semantic compositionality. In *Topoi*, volume 13, pages 11–24.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, Online. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, and Ethan Perez. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv.2307.11768*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv:2404.18796*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan

Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Jian Wu, Linyi Yang, Yuliang Ji, Wenhao Huang, Börje F. Karlsson, and Manabu Okumura. 2024. Gendec: A robust generative question-decomposition method for multi-hop reasoning. *arXiv:2402.11166*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. 2025. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3971–3992, Vienna, Austria. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Kun Zhang, Jiali Zeng, Fandong Meng, Yuanzhuo Wang, Shiqi Sun, Long Bai, Huawei Shen, and Jie Zhou. 2024. Tree-of-reasoning question decomposition for complex question answering with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19560–19568.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv:2303.18223*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Dataset Details

Table 4 shows the sample count per hop and total count for each dataset.

	Hotpot	2Wiki-complex	Strategy
2-hop	787	166	86
3-hop	-	347	208
4-hop	-	243	42
Total	787	756	336

Table 4: The number of samples per hop and the total number of samples for the three datasets.

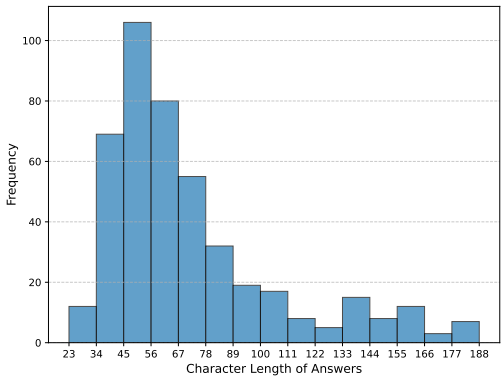

Figure 2: Histogram showing the character lengths of sub-answers.

A.1 2Wiki-complex

Specifically, (**Step 1**) from the development set of 2Wiki, we only retain the questions related to father or mother relations (1,025 samples). The primary reason for this is that the combination of the two parent relations can form a new relational word, such as 'mother' and 'mother' combining to create 'maternal grandmother'. (**Step 2**) Since 2Wiki is constructed automatically, it contains unanswerable questions. Therefore, we randomly selected 200 samples and manually verified them to ensure that all samples we use for generating more hops are answerable, resulting in 182 2-hop samples. (**Step 3 & 4**) We prepare a list of templates for extending questions from n-hop to $(n+1)$-hop, where n can be 2 or 3. Table 5 presents a list of templates that we use to extend from 2-hop to 3-hop. For example, if the 2-hop question is [*Who is the maternal grandmother of person A?*], the corresponding 3-hop question would be [*What is the date of birth of the maternal grandmother of person A?*]. After obtaining all necessary information, we automatically generate the samples, resulting in 392 3-hop samples, and 266 4-hop samples.

This is an example of the process for extending a 2-hop sample to a 3-hop sample. Each 2-hop sample corresponds to two triples, (e_1, r_1, e_2) and (e_2, r_2, e_3), and two paragraphs (p_1 and p_2) that describe the two entities, e_1 and e_2. To obtain a new 3-hop sample from a 2-hop sample, we start with the entity e_3 and obtain a list of new triples, such as (e_3, r_{31}, e_4) and (e_3, r_{32}, e_4). We manually verify this list of new triples. When a triple (e.g., (e_3, r_{31}, e_4)) is confirmed to be correct, we use the three triples (e_1, r_1, e_2), (e_2, r_2, e_3), and (e_3, r_{31}, e_4) to generate a 3-hop sample.

A.2 StrategyQA

Notably, unlike HotpotQA and 2Wiki, StrategyQA consists of yes/no questions. Geva et al. (2021) provided the following example to illustrate the difference between explicit and implicit multi-hop reasoning:

- **Explicit Question**: Was Aristotle alive when the laptop was invented?
- **Implicit Question**: Did Aristotle use a laptop?

Sample Selection. The training and test sets contain 2,290 and 490 samples, respectively. However, since sub-questions are unavailable in the test set, we select samples only from the training set. To consolidate evidence from different annotators, we collect all annotated paragraph titles and remove duplicates. This results in 87, 257, 443, 513, 428, 284, and 159 samples with 1 to 7 paragraphs, respectively, along with a few samples containing more than 7 paragraphs. Since samples with multiple paragraphs may contain multiple reasoning paths, we focus on those with 2 to 4 paragraphs. The training set contains 18, 626, 1,219, 342, and 85 samples with 1 to 5 sub-questions. To maintain a balanced level of complexity, we further restrict our selection to samples with 2 to 4 sub-questions.

Sub-Answer Generation Process. We run Llama-3.3 70B with few-shot prompting to generate sub-answers for the sub-questions. Since the sub-questions are dependent, we process them step by step to obtain their answers. Our prompting incorporates both provided short facts and context because short facts are more useful in helping the

Relation_1	Relation_2	Relation_3	Question_template
father	father	father	Who is the paternal grandfather's father of #name?
father	father	mother	Who is the paternal grandfather's mother of #name?
father	father	spouse	Who is the paternal grandfather's wife of #name?
father	father	date of birth	What is the date of birth of paternal grandfather of #name?
father	father	date of death	What is the date of death of paternal grandfather of #name?
father	father	place of birth	What is the place of birth of paternal grandfather of #name?
father	father	place of death	What is the place of death of paternal grandfather of #name?
father	father	country of citizenship	What is the nationality of paternal grandfather of #name?
father	father	cause of death	What is the cause of death of paternal grandfather of #name?
father	mother	father	Who is the paternal grandmother's father of #name?
father	mother	mother	Who is the paternal grandmother's mother of #name?
father	mother	spouse	Who is the paternal grandmother's husband of #name?
father	mother	date of birth	What is the date of birth of paternal grandmother of #name?
father	mother	date of death	What is the date of death of paternal grandmother of #name?
father	mother	place of birth	What is the place of birth of paternal grandmother of #name?
father	mother	place of death	What is the place of death of paternal grandmother of #name?
father	mother	country of citizenship	What is the nationality of paternal grandmother of #name?
father	mother	cause of death	What is the cause of death of paternal grandmother of #name?
mother	mother	father	Who is the maternal grandmother's father of #name?
mother	mother	mother	Who is the maternal grandmother's mother of #name?
mother	mother	spouse	Who is the maternal grandmother's husband of #name?
mother	mother	date of birth	What is the date of birth of maternal grandmother of #name?
mother	mother	date of death	What is the date of death of maternal grandmother of #name?
mother	mother	place of birth	What is the place of birth of maternal grandmother of #name?
mother	mother	place of death	What is the place of death of maternal grandmother of #name?
mother	mother	country of citizenship	What is the nationality of maternal grandmother of #name?
mother	mother	cause of death	What is the cause of death of maternal grandmother of #name?
mother	father	father	Who is the maternal grandfather's father of #name?
mother	father	mother	Who is the maternal grandfather's mother of #name?
mother	father	spouse	Who is the maternal grandfather's wife of #name?
mother	father	date of birth	What is the date of birth of maternal grandfather of #name?
mother	father	date of death	What is the date of death of maternal grandfather of #name?
mother	father	place of birth	What is the place of birth of maternal grandfather of #name?
mother	father	place of death	What is the place of death of maternal grandfather of #name?
mother	father	country of citizenship	What is the nationality of maternal grandfather of #name?
mother	father	cause of death	What is the cause of death of maternal grandfather of #name?

Table 5: List of question templates that are used to extend a question from 2-hop to 3-hop in the 2Wiki-complex dataset.

model generate accurate sub-answers. After obtaining all sub-answers, we identify 508 samples where the predicted final answer aligns with the gold final answer. However, we observe that some sub-answers are excessively long. Figure 2 presents a histogram of sub-answer character lengths. Since each question may have multiple sub-questions and sub-answers, we take the maximum sub-answer length within a sample to represent it. Based on this histogram and manual inspection, we retain only samples where sub-answers are under 100 characters.

B Experimental Results

B.1 Decomposition Stage

Human Evaluation Guideline. We present the following evaluation guidelines to the annotators.

- **Perfection Decomposition**: 1 and 0
  - All sub-questions are connected, and the final sub-question should lead to the correct final answer.
  - All sub-questions are useful (each contributing to the final answer) and not redundant.
  - If you label it as 1, no further annotations are needed for this sample.
- **Reasoning Chain**: 1 and 0
  - All sub-questions are connected, and the final sub-question should lead to the correct final answer.
  - The sub-question cannot be the same as the original question.
  - Redundancy or missing sub-questions are differences from the perfect decomposition score.
- **Number of Correct Sub-questions**: We present the annotators with a list of 0s corresponding to the number of generated sub-questions. If the annotators believe a sub-question is correct and useful for finding the final answer, they can change the label from 0 to 1 for that sub-question. The final score is calculated by dividing the number of correctly generated sub-questions by the total number of generated sub-questions.

It is important to note that the evaluation was conducted by two authors of the paper. We have discussed our guidelines, as well as any special cases that caused confusion or were unclear.

Dataset	Llama 3.1 8B		Llama 3.1 70B		o1	
	EM	F1	EM	F1	EM	F1
Hotpot	66.2	82.6	68.2	83.2	62.1	78.0
2Wiki-c	73.4	88.5	80.0	92.6	71.3	89.4
Strategy	44.5	55.0	46.8	57.1	17.9	30.7

Table 6: EM and F1 scores in the sub-problem-solving stage for the case where all sub-questions are of equal importance.

B.2 Subproblem Solving Stage

Table 6 presents EM and F1 scores in the sub-problem-solving stage for the case where all sub-questions are of equal importance.

We provide examples of the differences between EM/F1 scores and LLM-as-a-judge in Table 7.

B.3 Decomposition Error Analysis

Table 8 presents a detailed error analysis of the decomposition stage for Llama 3.1 70B and o1.

We provide error examples from the decomposition stage in Table 9.

Example	Question	Predicted Answer	Gold Answer	Scores
1	When was the poet Rumi active?	30 September 1207 – 17 December 1273	13th century	**EM**: false **F1**: 0.0 **LLM**: CORRECT
2	In a monopoly, how many different entities supply goods?	One	1	**EM**: false **F1**: 0.0 **LLM**: CORRECT
3	What class of animals do silverfish belong to?	Insects	Insect	**EM**: false **F1**: 0.0 **LLM**: CORRECT
4	What ages are most medicare recipients?	65 and older, as well as some younger people with disability status as ...	65 or older	**EM**: false **F1**: 0.15 **LLM**: CORRECT
5	What is the maximum depth of the Sea of Japan?	3,742 meters (12,277 ft)	12,276 feet (3,742 metres)	**EM**: false **F1**: 0.25 **LLM**: CORRECT
6	What is the range of a Hwasong-15 missile?	More than 13,000 km (8,100 miles)	8,000 miles	**EM**: false **F1**: 0.25 **LLM**: CORRECT
7	Who is the father of Princess Victoria Melita of Saxe-Coburg and Gotha	Alfred, Duke of Saxe-Coburg and Gotha	Alfred	**EM**: false **F1**: 0.28 **LLM**: CORRECT
8	How big is Walt Disney World in square miles?	39 sq mi	39	**EM**: false **F1**: 0.5 **LLM**: CORRECT
9	Who is the father of Princess Catherine Of Schleswig-Holstein-Sonderburg-Beck?	Duke Peter August of Holstein-Beck	Peter August	**EM**: false **F1**: 0.57 **LLM**: CORRECT
10	When was the play Dido, Queen of Carthage written?	It was probably written between 1587 and 1593	between 1587 and 1593	**EM**: false **F1**: 0.67 **LLM**: CORRECT
11	Worldview Entertainment is an American independent motion picture company that produced which 2015 mystery thriller film based on a novel?	Child 44	Child 44 (film)	**EM**: false **F1**: 0.8 **LLM**: CORRECT
12	Who is the father of Godfrey I, Count of Verdun?	Gozlin, Count of Bidgau and Methingau	Gozlin, Count of Bidgau	**EM**: false **F1**: 0.8 **LLM**: CORRECT

Table 7: Examples of differences between EM/F1 scores and LLM-as-a-judge.

Dataset	Llama 3.1 70B			o1		
	#Err	#Ref	#Acc	#Err	#Ref	#Acc
Hotpot (95)	40	2	0	35	23	15
2Wiki-c (100)	15	0	0	13	1	1
Strategy (92)	54	1	0	74	45	12

Table 8: Detailed error analysis in the decomposition stage: #Err represents the number of decomposition error cases, #Ref indicates the number of cases related to referencing issues, and #Acc counts the number of cases that become acceptable once the referencing issue is resolved.

Dataset	Question	Generated Decomposition	Gold Decomposition	Error Type
Strategy	Does the anatomy of a camel lend itself to jokes on Wednesdays?	1. Which notable aspect of this animal's anatomy might be relevant to lighthearted remarks? 2. Does #Ans1 share a term or concept with a mid-week day in typical work schedules? 3. If #Ans2 is yes, does that shared reference inspire humor on that day?	1. As a joke, what is Wednesday otherwise known as? 2. What are camels known for having? 3. Is there overlap between #1 and #2?	Referencing issue *Accept after fixing*
Strategy	Are ground bell peppers the main ingredient of black pepper?	1. Do the two items in the question come from the same plant source? 2. If #Ans1 is no, can ground bell peppers be the main ingredient of black pepper?	1. What is used to make black pepper? 2. Is #1 the same thing as bell pepper?	Referencing issue *Do not accept*
Strategy	Was Rumi's work serialized in a magazine?	1. Did this individual produce any written works? 2. Were any of #Ans1 published in multiple parts over time? 3. If #Ans2 is yes, did those publications appear in a magazine format?	1. When was the first magazine ever published? 2. When was the poet Rumi active? 3. Was #1 before #2?	Referencing issue *Do not accept*
Hotpot	When did the Disney sitcom in which Genevieve Knight G̈Ḧannelius starred as Avery Jennings end?	1. Which TV show featured this actress in the role of Avery Jennings? 2. When did #Ans1 end?	1. What is the Disney sitcom that Genevieve Knight G̈Ḧannelius starred as Avery Jennings? 2. When did #1 end?	Referencing issue *Accept after fixing*
Hotpot	The director of a film about Wong Fei-hung also made what film?	1. Which motion picture focuses on the Chinese folk hero referenced in the original inquiry? 2. Who directed #Ans1? 3. Which other film did #Ans2 also create?	1. Who is the director of a film about Wong Fei-hung? 2. #1 also made what film?	Referencing issue *Do not accept*

Table 9: Examples of errors in the decomposition stage. *Accept after fixing* means it is acceptable once the referencing issue is fixed. *Do not accept* means it is not acceptable even if the referencing issue is fixed.