# MoreHopQA: More Than Multi-hop Reasoning

**Anonymous ACL submission**

## Abstract

Most existing multi-hop datasets are extractive answer datasets, where the answers to the questions can be extracted directly from the provided context. This often leads models to use heuristics or shortcuts instead of performing true multi-hop reasoning. In this paper, we propose a new multi-hop dataset, `MoreHopQA`, which shifts from extractive to generative answers. Our dataset is created by utilizing three existing multi-hop datasets: HotpotQA, 2WikiMultihopQA, and MuSiQue. Instead of relying solely on factual reasoning, we enhance the existing multi-hop questions by adding another layer of questioning that involves one, two, or all three of the following types of reasoning: commonsense, arithmetic, and symbolic. Our dataset is created through a semi-automated process, resulting in a dataset with 1,118 samples that have undergone human verification. We then use our dataset to evaluate five different large language models: Mistral 7B, Gemma 7B, Llama 3 (8B and 70B), and GPT-4. We also design various cases to analyze the reasoning steps in the question-answering process. Our results show that models perform well on initial multi-hop questions but struggle with our extended questions, indicating that our dataset is more challenging than previous ones. Our analysis of question decomposition reveals that although models can correctly answer questions, only a portion—38.7% for GPT-4 and 33.4% for Llama3-70B—achieve perfect reasoning, where all corresponding sub-questions are answered correctly.[1]

**An Existing Multi-hop Sample**

**Question:** What is the date of birth of the father of Louis XIV?

***Paragraph A: Louis XIV***

Louis XIV (5 September 1638 – 1 September 1715), also known as Louis the Great … Louis XIV was born on 5 September 1638 in the Château de Saint-Germain-en-Laye, to Louis XIII and …

***Paragraph B: Louis XIII***

Louis XIII (27 September 1601 – 14 May 1643) was King of France from 1610 until his death in 1643 and King of Navarre (as Louis II) from 1610 to 1620, …

**Answer:** 27 September 1601

**A New Sample in Our Dataset**

**New Question:** What is the date 3 weeks after the date of birth of the father of Louis XIV?

**New Answer:** October 18, 1601

**Question Decomposition:**

**Sub-question 1:** Who is the father of Louis XIV?

**Sub-question 2:** What is the date of birth of Sub-ans-1?

**Sub-question 3:** What is the date 3 weeks after Sub-ans-2?

Figure 1: An example of our dataset. Our new question is created by extending the initial 2-hop question, which ensures that the new answer is generative.

## 1 Introduction

Multi-hop Question Answering (QA) requires a model to retrieve, extract, and connect pieces of evidence from multiple paragraphs to answer a question (Welbl et al., 2018; Yang et al., 2018). By harnessing the reasoning abilities of models, this task provides valuable insights into evaluating their capabilities in understanding natural language and tackling complex tasks. For this reason, multi-hop QA has received much attention over the past few years, prompting the creation of several benchmark datasets such as HotpotQA (Yang et al., 2018), 2WikiMultihopQA (2Wiki; Ho et al., 2020), MuSiQue (Trivedi et al., 2022), MQuAKE (Zhong et al., 2023), MRKE (Wu et al., 2024), or FanOutQA (Zhu et al., 2024).

While existing multi-hop QA datasets have been instrumental in evaluating the reasoning capabilities of Large Language Models (LLMs), they suffer from several limitations. The first limitation concerns the type of answers found in these datasets. Indeed, most of the answers are extractive, meaning they can be directly extracted from the supporting

[1]We will release our data and code in the future. For this submission, we have included them in the zip file.

paragraphs provided as context. Such answers may incentivize models to generate answers through heuristics or reasoning shortcuts (Min et al., 2019a; Geirhos et al., 2020; Ho et al., 2023), rather than engaging in the expected multi-step reasoning task. For example, questions asking about dates with supporting paragraphs containing only one possible date entity are likely to be guessed correctly by models. The second limitation lies in the restricted range of reasoning types found in existing multi-hop datasets, which primarily focus on reasoning tasks involving common knowledge from Wikipedia. Consequently, they neglect other forms of reasoning, such as arithmetic or symbolic reasoning, which are also crucial to consider when evaluating the reasoning capabilities of models (Qiao et al., 2023).

In this paper, we aim to address these limitations by introducing MoreHopQA, a new dataset made of multi-hop questions whose answers cannot be simply extracted and instead require combining multiple types of reasoning. Our approach involves extending questions from existing datasets with additional hops, thereby transforming their original answers into generative answers, which prevents them from being simply guessed by models (see Figure 1). More specifically, our dataset features the following main aspects: 1) Answers are generative, requiring models to reason to derive the final answer. 2) To answer questions in our dataset, models need to engage in multi-step reasoning first, followed by another type of reasoning (e.g., arithmetic). 3) We provide explicit decompositions, that is, the set of sub-questions and sub-answers in the reasoning process from question to answer. We argue that adopting generative answers and challenging models to perform additional types of reasoning beyond multi-hop questions can make the dataset more demanding for the models.

Our dataset creation process involves the following four steps: 1) *Sample Selection* (§3.1), where we manually curated 2-hop samples from three existing multi-hop datasets (i.e. HotpotQA, 2Wiki, and MuSiQue) according to three criteria: questions should be answerable, include sub-questions and sub-answers, and have properly formatted answers. 2) *Template Design* (§3.2), where we (the authors of this paper) collaboratively designed about 100 templates for creating new questions encompassing three types of reasoning (i.e. arithmetic, commonsense, and symbolic) from five answer types (i.e. person, place, organization, date and year). 3) *New Sample Generation* (§3.3), where we use our templates in conjunction with the selected 2-hop samples to automatically generate new samples. 4) *Human Verification* (§3.4), where we ensure the quality of our new samples by asking a pool of annotators to label and revise them, resulting in a final dataset of 1,118 human verified samples. We further validate the quality of our dataset by evaluating human performance on a subset of 150 samples, demonstrating that our new samples are both answerable and reasonable (§4).

We then use our dataset to evaluate the reasoning capabilities of five different LLMs: Mistral 7B, Gemma 7B, Llama 3 (8B and 70B), and GPT-4. We conduct experiments using multiple prompting strategies, including zero-shot, few-shot, and Chain-of-Thought (CoT) (Wei et al., 2022). We leverage the explicit decompositions of the questions in our dataset to conduct an extensive error analysis (Figure 2), precisely identifying where in the reasoning chain the models fail and highlighting which models resort to reasoning shortcuts. Our results indicate that while the models perform well on the initial multi-hop questions, they struggle more with our extended questions. This suggests that our dataset presents a greater challenge compared to previous datasets. Our analysis of question decomposition reveals that while models can correctly answer questions, only a small portion (38.7% for GPT-4 and 33.4% for Llama3-70B) achieve perfect reasoning, where all corresponding sub-questions are answered correctly.

In summary, our contributions are as follows:

- We create a more challenging dataset that shifts from extractive to generative, and, with the decompositions, allows for a better understanding of the reasoning capabilities of LLMs.
- We conduct extensive human verification and validation to ensure the quality of our dataset.
- We evaluate the performance of five LLMs and show that even state-of-the-art LLMs do not match human performance. We also find that while GPT-4 performs best, only 38.7% reach the state of perfect reasoning.

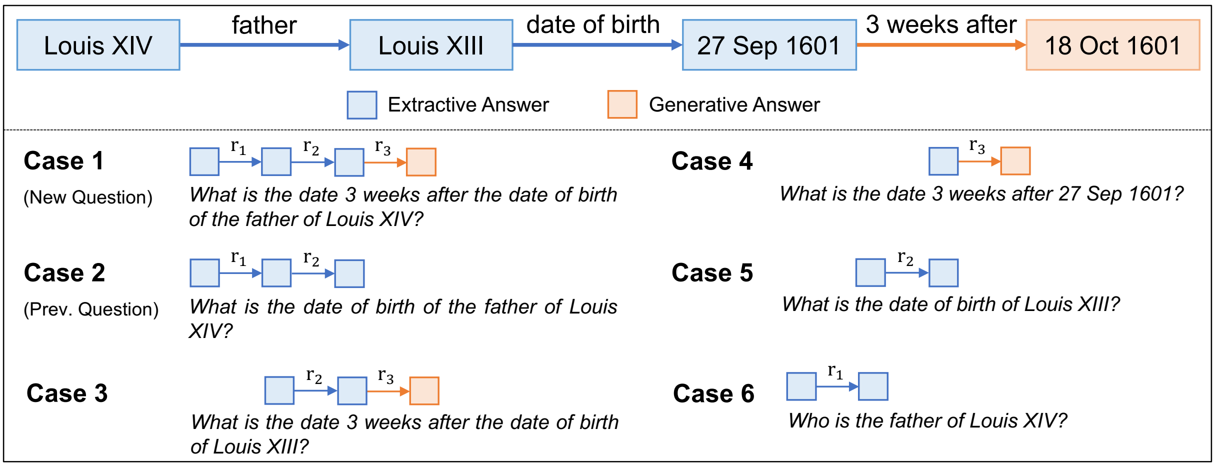


Figure 2: There are six cases in our analyses. The first case is our newly generated question. The second case is the initial 2-hop question. We present the details of these cases in Appendix A.6.

## 2 Related Work

### 2.1 Multi-hop QA Datasets

The first multi-hop QA dataset, QAngaroo, was introduced by Welbl et al. (2018). It consists of two sub-datasets, WikiHop and MedHop, and was constructed by leveraging both unstructured text sources (e.g. Wikipedia or Medline) and structured data from external resources (e.g. Wikidata or DrugBank). In the same year, Talmor and Berant (2018) introduced ComplexWebQuestions, a dataset derived from WebQuestionsSP (Yih et al., 2016) that contains automatically generated questions revised by crowdworkers. In the following years, HotpotQA (Yang et al., 2018), R$^4$C (Inoue et al., 2020), 2WikiMultihopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022) were introduced, with a greater emphasis on explaining the QA process. MQuAKe (Zhong et al., 2023) and FanOutQA (Zhu et al., 2024) are two recently proposed datasets. MQuAKe focuses on testing multi-hop reasoning for knowledge editing in LLMs, while FanOutQA focuses on creating complex listing questions. However, many existing datasets only feature extractive answers and focus solely on multi-hop reasoning within Wikipedia text. In contrast, our dataset shifts from extractive to generative answers, requiring broader reasoning abilities for answering the questions.

### 2.2 Multi-hop Analyses

Due to the intricate nature of multi-hop questions, they are particularly useful for analyzing and evaluating the reasoning chains in the QA process. Tang et al. (2021) utilized sub-questions in the QA process and conducted experiments on HotpotQA to determine whether multi-hop models could answer them successfully. They found that multi-hop models did not perform well on this task.

Trivedi et al. (2020) used the connection between the two supporting facts to analyze the abilities of the models. They found that even with disconnections, the models could still answer the questions, revealing that the models can use heuristics or shortcuts to arrive at the answers. In the shortcuts analyses, several previous works (Min et al., 2019a; Chen and Durrett, 2019; Jiang and Bansal, 2019) also raised the issues about the multi-hop reasoning abilities of the models and the shortcuts in existing datasets.

Additionally, recent works (Dua et al., 2022; Khot et al., 2023; Press et al., 2023; Zhou et al., 2023) attempted to incorporate a question decomposition step into their prompts to improve model performance. Prior to these studies, some works (Talmor and Berant, 2018; Min et al., 2019b; Fu et al., 2021) showed that integrating question decomposition into their systems can lead to better performance and more explainable responses. Patel et al. (2022) showed that human decomposition improves performance on complex questions. However, Wei et al. (2023) showed that question decomposition does not help when there are more samples in the dataset. Due to sparse benchmarks, drawing reliable conclusions about question decomposition is challenging. Our dataset includes sub-questions and sub-answers, which could be valuable for future research on exploring the effectiveness of question decomposition.

## 3 Dataset Creation Process

Our dataset creation process, illustrated in Figure 3, consists of four main steps: 1) sample selection, 2) template design, 3) new sample generation, and 4) human verification. We first describe each of these four steps and then provide detailed information about the final version of the dataset.

### 3.1 Sample Selection

Our new samples are derived from 2-hop questions found in three existing multi-hop datasets: HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). To ensure the quality of our dataset, we defined three criteria for selecting the initial 2-hop samples: 1) **Answerability**: all 2-hop questions should be answerable, that is, the answer must be found in the supporting paragraphs. 2) **Decomposition**: initial 2-hop samples should contain a list of sub-questions and sub-answers. 3) **Format**: we categorized the initial 2-hop samples based on their answer type, such as person name, date, year, or location, and applied specific requirements to each group. For example, dates should be fully formatted (comprising day, month, and year), while person names should include both the first and last names. Herein, we describe the methodology we applied for selecting the initial samples from each dataset.

**HotpotQA** Since the original HotpotQA lacks sub-questions and sub-answers, we relied on Tang et al. (2021), who annotated 1,000 samples with them. From this pool, we manually curated a subset of samples, discarding those that are difficult to understand or have answers in an incorrect format, and annotated each sample with its corresponding answer type. Notably, we observed that the format of the answers for the place type was inconsistent, making it difficult to integrate with templates, so we decided to exclude them. We obtained 48, 47, and 19 samples with answer types of person, year, and date, respectively.

**2Wiki** We selected the bridge questions from the development set as our initial samples. Based on the relation type of the second triple in the reasoning chain, we classified the samples into five answer types: place, person, year, date, and string. Since questions in 2Wiki are automatically generated, we manually reviewed 400 samples to check their answerability and decide whether to use them. For instance, we opted to exclude questions with answers of the string type, as they often have multiple valid answers. We obtained 120, 114, 69, and 11 samples for place, date, person, and year, respectively.

**MuSiQue** We selected the samples from the development set with a structured format (similar to a triple format) for the second hop in the question decomposition process. Based on the relation information of the second hop, we automatically annotated the answer types of the samples, resulting in 105, 99, and 22 samples for person, place, and organization, respectively. We observed that a substantial number of samples in MuSiQue have multiple answers, being either explicitly indicated in the dataset (`answer_aliases` field) or identified during our manual verification process. Because our new answers are based on the answers to the 2-hop questions, we do not include these samples in our dataset. As a result, we obtain 17, 14, and 3 samples for person, place, and organization, respectively. We present examples of these issues in Appendix A.2, which further explains why the final number of samples drawn from MuSiQue is small.

### 3.2 Template Design

We, the authors of this paper, collaboratively designed 97 templates for creating the new questions in our dataset. Multiple templates were designed for each answer type, with the purpose of creating new questions whose new answers are generative, meaning they can not be simply extracted from the supporting paragraphs. For example, regarding the date answer type, we can ask about the next day, next month, next week, next year, or any other gap relative to the current date. Another example for the person name answer type, we can ask about the first letter of the first name, the last letter of the first name, or the concatenation of the first letter and last letter of the first name. As discussed in the Introduction, we conjecture that extractive answers are easy for models to identify, potentially leading to their tendency to rely on heuristics and shortcuts in the QA process. Here, we purposely crafted our templates to address that issue, adding one extra hop to the initial 2-hop question to make the new answer a generative type.

In Qiao et al. (2023), five types of reasoning are explored: arithmetic reasoning, commonsense reasoning, symbolic reasoning, logical reasoning, and multimodal reasoning. We designed our templates to encompass the first three types of reasoning, but

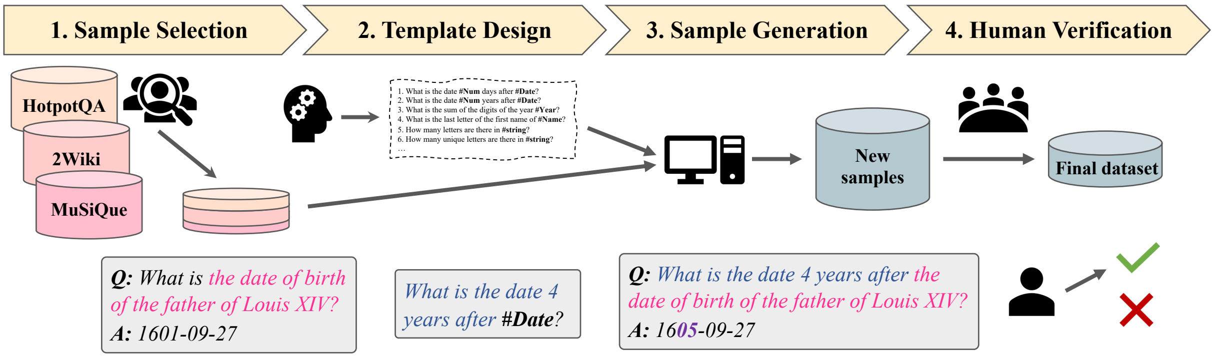


Figure 3: Our dataset creation process.

not extend to logical or multimodal reasoning due to the nature of the samples we use (multi-hop questions in the Wikipedia domain). Our templates cover all three of these reasoning types individually, as well as various combinations thereof. Some templates rely on a single type of reasoning, while others require two or three types. Each template is labeled with its corresponding reasoning type(s), and we also indicate the number of hops required to answer the new questions. If the number of required hops exceeds one, we include a list of sub-questions and their corresponding sub-answers.

### 3.3 New Sample Generation

We use the list of templates in conjunction with the selected 2-hop samples to generate new samples for our dataset. This involves creating both a new question and a new answer for each pair of template and 2-hop sample. To generate a new question, we combine our templates with the noun phrases extracted from the initial 2-hop questions. For example, given the question [ *What is the date of birth of the father of Louis XIV?* ] and our template [ *What is the date one week after #Date?* ], we first extract the noun phrase of the question [ *the date of birth of the father of Louis XIV* ]. Next, we replace the special token *#Date* in our template by this noun phrase to get [ *What is the date one week after the date of birth of the father of Louis XIV?* ]. We also incorporate another special token *#Num* for numerical quantities, allowing us to choose various values (e.g., one week, two weeks) when generating new questions. The 2-hop questions in 2Wiki and MuSiQue are well-structured, allowing us to extract their noun phrases using rule-based methods. However, as the HotpotQA questions are crowdsourced, we resort to manual annotation to accurately identify the noun phrase of each question. To obtain the new answer, we use code to perform the operations on the initial 2-hop answer corresponding to the template (e.g., adding one week). An example of a generated sample is provided in Appendix A.5.

From 114, 314, and 34 samples in HotpotQA, 2Wiki, and MuSiQue, respectively, we generate 1,497, 2,617, and 373 new samples. There are four answer types in our dataset: date, number, string, and letter. Statistics about the number of samples for each type are presented in Table 1. An example question for each answer type is provided in Appendix A.6.

### 3.4 Human Verification

After completing the previous steps, we have generated a total of 4,487 new samples. Our focus now shifts to ensuring the quality of our dataset, as these newly generated questions may exhibit issues stemming from our template-based approach. We extracted a subset of 1,408 randomly selected new samples for human verification and tasked 10 annotators (students and researchers in NLP, including the authors) with verifying and, if necessary, modifying the generated questions. The human verification process involves labeling the new questions with one of the following three labels: [**OK**] the question is acceptable and requires no changes; [**Modified**] the question had flaws that were corrected through modifications; and [**Issue**] the question has significant problems that remain despite attempts to modify it. The guidelines and the annotation interface are provided in Appendix A.4. Out of the 1,408 samples that were verified, 919 were labeled as OK, 408 as Modified, and 81 as Issue. Questions labeled as Issues were double-checked, and those deemed unusable (e.g., initial 2-hop question having multiple answers) were discarded from our final dataset.

| Dataset | Date | Number | String | Letter | Total |
|---|---|---|---|---|---|
| HotpotQA | 76 | 1,070 | 304 | 47 | 1,497 |
| 2Wiki | 567 | 1,453 | 528 | 69 | 2,617 |
| MuSiQue | 17 | 225 | 114 | 17 | 373 |
| MoreHopQA w/ hv | 216 | 663 | 196 | 43 | 1,118 |
| MoreHopQA w/o hv | 436 | 1,526 | 479 | 61 | 2,502 |

Table 1: Statistics showing the number of generated samples for each answer type in our dataset. MoreHopQA w/ hv indicates the version with human verification.

### 3.5 Final Dataset

After the human verification process, we are left with 1,118 new samples. Statistics for the number of samples for each answer type in our final dataset are presented in Table 1. In addition to the subset that underwent human verification, we also release the remaining subset of 2,502 samples without human verification. For this latter subset, we automatically filtered out the samples derived from questions marked as erroneous through the human verification process, aiming to enhance its overall quality. Our dataset information is in English.

## 4 Dataset Quality Assessment

To further validate the quality of our dataset and provide an estimate of human performance, we tasked the same pool of annotators as in §3.4 with answering a randomly selected subset of 150 samples. Each sample consists of a question and two supporting (gold) paragraphs. The task of the annotators is to answer the given questions. Each sample is annotated by two separate annotators. Since our aim is to assess the reasoning abilities of the process rather than focusing on its retrieval components, we do not include distractor paragraphs.

We calculate three distinct metrics: the average human performance, the human upper bound, and the inter-annotator agreement. Following (Yang et al., 2018; Ho et al., 2020), the upper bound is computed as the average of maximum exact match (EM) for each sample. We obtain scores of 84.3, 94.0, and 76.7 for these three metrics, respectively. The notably high human performance scores, encompassing both the average and upper bound, serve as strong indicators of the quality of the dataset. Notably, the human performance average score sets a benchmark for the expected model performance. Furthermore, the inter-annotator agreement score, although slightly lower, remains within an acceptable range, affirming the consistency and reliability of our dataset.

## 5 Experiments

### 5.1 Experimental Settings

**Models** We compare the performance of several instruction-fine-tuned auto-regressive LLMs on our dataset. To represent a variety of current models in terms of size and fine-tuning, we chose Llama-3-8B-Instruct and Llama-3-70B-Instruct from the Llama-3 family of models (AI@Meta, 2024), as well as Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Gemma-7B (Team et al., 2024), and GPT-4 Turbo (OpenAI et al., 2024).

**Prompting** Following the results from Kojima et al. (2022) and Wei et al. (2022), we compare the performance using zero-shot and few-shot prompting with 2 and 3 shots, as well as CoT prompting with zero, 2, and 3 shots. For comparability, we use the same user prompts for all models. The only variation in our prompting setup is the inclusion of a system prompt, which is applied when specified by the model's authors in its Hugging Face model card. We select the few-shot examples from our dataset in such a way that the answer types of the examples match those of our question, while ensuring that none of the answers to the subquestions are revealed in the prompt.

**Baseline** Following previous work on detecting potential reasoning shortcuts in datasets (Sugawara et al., 2018; Trivedi et al., 2022), we run an artifact-based baseline with Llama-8B. In this baseline, we only use the two words from the question (e.g., “when was” or “how many”).

**Evaluation** We follow the general approach of evaluating multi-hop QA tasks as presented in (Yang et al., 2018), and additionally run postprocessing on the generated model output to extract the final answer, depending on the expected type of the answer. When prompting, we ask the model to give the final answer between two <answer> tags, and parse the string between those as the model's final answer. We then attempt to convert this string into the respective built-in python datatype for the answer type, either directly or with the help of Named Entity Recognition, and convert it back to a default string representation. We then report the EM and F1 scores on the tokens between the preprocessed ground-truth answer and the postprocessed model-generated answer.

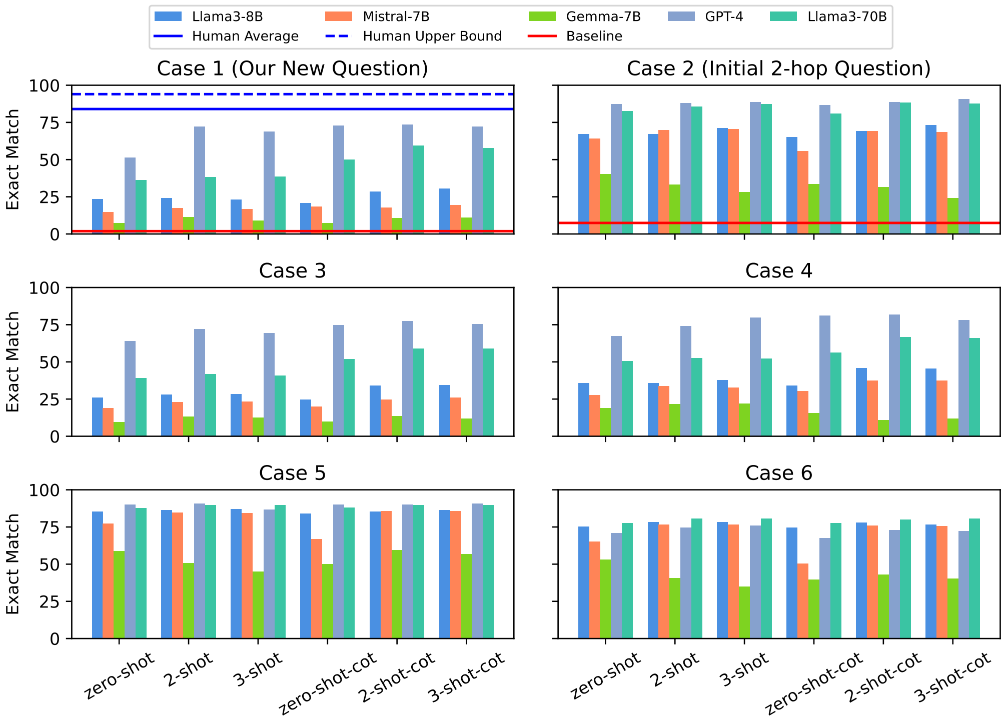


Figure 4: Performance (EM scores) of the models on our dataset.

## 5.2 Results

The performance (EM scores) of all models on our dataset are presented in Figure 4. We present both EM and F1 scores in Appendix B.2.

**Baseline Performance** We observe that the performance of the baseline is low but non-zero, and better on the initial 2-hop questions (1.9 EM and 7.4 EM). As the scores are far from any other model's performance in both cases, this indicates that the models cannot directly use heuristics to solve most questions.

**Models vs. Human Performance** As shown in §4, the average human performance and the human upper bound are 84.3 and 94.0, respectively. However, even in the best setting, GPT-4's performance is still lower than the average human score, indicating that there is room for improving the reasoning abilities of current models.

**Our Question vs. Initial 2-hop Question** Between the initial 2-hop questions (Case 2) and our new questions (Case 1), we observe a decrease in performance for both EM and F1 scores across all models when adding an additional hop, between up to 26.0 points in EM for GPT-4, to up to 53.8 points EM for Mistral-7B. Smaller models such as Mistral-7B and Llama3-8B seem to have a larger gap in performance between both cases compared to larger models. This indicates that our dataset is more challenging than the initial 2-hop datasets.

**CoT Prompting** All tested models benefit from the few shot-CoT prompting, gaining between 3.5 (Mistral-7B) and 23.0 (GPT-4) percentage points EM. The best performance is reached by GPT-4_2-shot-cot prompting, which reaches 73.3 EM. Generally, larger models perform better, as both GPT-4 and Llama-70B reaching up to 73.3 and 59.2 EM, respectively, compared to between up to 11.3 and 30.5 EM for the models with 7-8 B parameters. During analysis, we observed that the result of Gemma-7B often refuses to answer. In our final results, we found from a total of 6,708 prompts, the answer contained the string "I cannot answer" up to 1,452 times (reached for 3-shot-cot).

**Results on Six Cases** As shown in the Figure, all models obtain high scores on the initial two-hop questions and its sub-questions (Case 2,5,6), but low scores on questions that include our added reasoning step (Case 1,3,4). It seems that our additional hop adds additional difficulty to the questions, apart from the fact that the questions get longer, since all models achieve higher scores on Case 5 and 6 compared to Case 4. We believe this is mainly due to the extractive answer type in Case 5 and 6. Similarly, when comparing Case 2 and Case 3, the models also achieve higher scores on Case

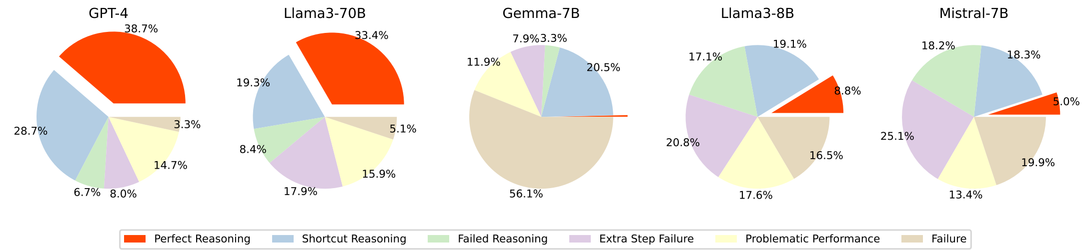


Figure 5: Distribution of performance categories of different LLMs on our dataset.

2 than on Case 3. In summary, our extended-hop approach increases the difficulty of the questions compared to the 2-hop extractive questions alone.

### 5.3 Performance Category Analysis

For a more detailed analysis of LLMs' performance, particularly the causes for the failures, we also ask the LLMs to answer the four other cases of the question, as shown in Figure 2. We classify LLMs' performance into the 6 following categories based on whether they can correctly answer different cases. We also present the detailed categorization in Appendix B.3.

- **Perfect Reasoning**: the LLM answers all cases correctly.
- **Shortcut Reasoning**: the LLM answers the initial question correctly, but fails in either of its sub-questions. In this situation, it extracts the answer from the context instead of reasoning.
- **Failed Reasoning**: the LLM answers the sub-questions correctly but fails in the question.
- **Extra Step Failure**: the LLM fails to answer all the cases regarding our designed question from the template. In this situation, it is unable to perform the required type of reasoning.
- **Problematic Performance**: the LLM answers the question correctly but inexplicably fails in some sub-questions, except shortcut reasoning.
- **Failure**: other conditions.

Figure 5 shows the distribution of performance categories of the LLMs on our dataset. All the models are prompted with 2-shot CoT examples because it shows the best overall performance across different models and cases. EM is the criterion used to determine whether the answer is correct or not. Consistent with the previous analysis, larger models (Llama3-70B and GPT-4) demonstrate more perfect reasoning compared with smaller models (Gemma-7B, Llama-8B, and Mistral-7B).

Llama3-7B and GPT-4 exhibit different performance patterns. Only 8% of extra step failure indicates that GPT-4 can better solve our designed template questions (Case 4) and their derivatives (Case 1, 3). For example, GPT-4 can correctly answer most questions in the format of *How many repeated letters are there in the first name of #Name?*, while Llama3-70B fails in some of these questions. It turns out Llama3 does not conduct arithmetic reasoning, commonsense reasoning and symbolic reasoning so well as GPT-4.

However, GPT-4 faces a substantial issue with shortcut reasoning. In 28.7% of the questions, GPT-4 can correctly answer the initial 2-hop question (Case 2) but fails in either of its sub-questions (Case 5 and Case 6). In contrast, Llama3-70B shows a "Shortcut Reasoning" rate of 19.3%. Thus, despite GPT-4's strong overall performance, our findings suggest that it heavily relies on shortcut reasoning to answer multi-hop questions. This highlights the need for a more detailed analysis when comparing the reasoning capabilities of different models

## 6 Conclusion

We introduce a new multi-hop dataset by extending existing 2-hop datasets with an additional hop. A notable aspect is that, through careful template design and selection of 2-hop samples, we transition from extractive to generative answers. Additionally, our samples require various types of reasoning to address the questions. Human performance scores indicate that our dataset is of high quality and suitable for evaluating models. We then use our dataset to evaluate the reasoning capabilities of five LLMs. Experimental results reveal a large gap between LLMs and human performance. Our analyses further demonstrate that the generative questions in our dataset are challenging for the models, preventing them from relying on simple heuristics to extract answers from the provided paragraphs.

## Ethical Statement and Broader Impact

Our dataset builds upon publicly available datasets, which themselves use publicly available information. The users were not asked to provide any information, and explicitly asked the users to fulfill a very narrow task, that did especially involve using only the available information. Human annotators were volunteer students on the Master's and PhD levels and professors working on research in an NLP Lab, who were given the opportunity to propose and execute their own annotation task with the same group of annotators in return. The annotators received an in-depth introduction including the topic of the research, and details about the intended use of the dataset.

Our work could help the community to benchmark new models and understand whether models are able to perform reasoning, an important next step in the development of intelligent models.

## Limitations

There are three limitations in our study. The first one concerns the diversity of the dataset. Although we try to use the three existing multi-hop datasets, our extended-hop questions are derived from designed templates (about 97 templates), which are not as diverse as non-template questions. The second point concerns our generated answers. These answers are not fully verified, as they are produced via code, based on the initial 2-hop answers. While we manually check the answers for all templates, we only verify a few samples per template, meaning not all answers are thoroughly reviewed. If unexpected cases occur that are not handled by our code, this may result in incorrect answers. The third point concerns running GPT-4. We have 6 settings per model, each with 6 cases (different types of questions), resulting in 36 runs per sample for one model. Due to the cost, we only ran GPT-4 on 150 samples.

## References


AI@Meta. 2024. Llama 3 model card.

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.

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.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, pages 665–673.

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.

Xanh Ho, Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2023. A survey on measuring and mitigating reasoning shortcuts in machine reading comprehension. *arXiv:2209.01824*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. 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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Is a question decomposition unit all we need? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4553–4569, Abu Dhabi, United Arab Emirates. 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.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, and et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

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.

Kangda Wei, Dawn Lawrie, Benjamin Van Durme, Yunmo Chen, and Orion Weller. 2023. When do decompositions help for machine reading? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3599–3606, Singapore. Association for Computational Linguistics.

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.

Jian Wu, Linyi Yang, Manabu Okumura, and Yue Zhang. 2024. MRKE: The multi-hop reasoning evaluation of llms by knowledge edition. *arXiv:2402.11924*.

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.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

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*.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. FanOutQA: Multi-hop, multi-document question answering for large language models. *arXiv:2402.14116*.

# A Dataset Creation Process

## A.1 Licenses

HotpotQA and MusiQue were published under the CC BY-SA 4.0 license, which explicitly allows adaptation. 2WikiMultihopQA was published under the Apache License 2.0, which also allows for distribution and modification. We intend to publish our newly generated dataset under the CC BY-SA 4.0 license.

## A.2 MuSiQue Dataset

We present three examples: (1) issues with disconnected reasoning, (2) lack of evidence to support the answer, and (3) multiple answers arising from setting questions without using the provided paragraphs in Tables 2, 3, and 4, respectively.

## A.3 Dataset generation details

We make use of various libraries to generate the answers to our dataset. For questions regarding the number of syllables, we make use of NLTK and use cmudict to estimate this number. To deal with place answers, we use the Nominatim API to search for places on OpenStreetView and retrieve the coordinates for each place mentioned in earlier datasets.

## A.4 Human Verification

We provide the following guidelines to annotators during the annotation process.

- Check the questions with *New Question (Overall)* or *New Question (Sub-question)* labels.
- If a question is good, give it an [**OK**] label.
- If a question is understandable but has some flaws (e.g., grammar, typo, etc.), give it a [**Modified**] label and please correct it.
- If a question is not understandable at all, give it an [**Issue**] label and briefly explain which part is confusing in the comment cell.
- Three additional fields are provided as **Reference**: *New Answer*, *Original Question*, and *Original Answer*. You don't need to check the correctness. However, if you find any severe issues (e.g., difficult to understand, the answer doesn't address the question, or messy code), please add a comment in the corresponding rows.

Figure 6 shows our annotation interface. We also provide the explanations for each field in the annotation guideline:

- *New Question (Overall)*: our new question
- *New Question (Sub-question)*: our new question but we only put the top question on the second hop. (in *New Question (Overall)*, we put the top question on the full 2-hop question)
- *New Answer*: an answer for a New Question (Overall)
- *Original Question*: the initial 2-hop question
- *Original Answer*: the answer for the Original Question

## A.5 Our Dataset Information

Each sample in our dataset contains the following information:

- _id: a unique id for each sample
- question: our new question
- answer: our new answer
- previous_question: the previous 2-hop question
- previous_answer: the previous 2-hop answer

**id:** 2hop__752214_639679
**question:** Who is the spouse of the author of Queen of the Elephants?
**answer:** Clio Goldsmith
**question_decomposition:**
- sub question 1: Queen of the Elephants » author
- sub answer 1: Mark Shand
- sub paragraph_support_title 1: Queen of the Elephants
- sub question 2: #1 » spouse
- sub answer 2: Clio Goldsmith
- sub paragraph_support_title 2: Clio Goldsmith

**context:**
*Paragraph 1: Queen of the Elephants*
Queen of the Elephants is a book written by the conservationist and travel writer Mark Shand and the corresponding BBC documentary Öueen of the Elephants, based on the life of the first female mahout in recent times–Parbati Barua of Kaziranga. The book went on to win the award, providing free publicity simultaneously to the profession of mahouts, and to Kaziranga.
*Paragraph 2: Clio Goldsmith*
Clio Goldsmith (born 16 June 1957) is a French former actress, appearing mostly as a Femme fatale in some films of the early 1980s. She is a member of the prominent Goldsmith family through her father ecologist Edward Goldsmith.

Table 2: This is an example of disconnected reasoning in MuSiQue: as shown in this example, from the answer of the first sub-question (Mark Shand), we have no evidence to proceed to the final answer (Clio Goldsmith).

| id | Information | Content | Status | Comment | _id |
|---|---|---|---|---|---|
| 1 | New Question (Overall) | What is the binary code of the first letter of the firstname of the author of the play that was adapted into a film and featured the orchestral arrangement Suite from Henry V in lowercase? | OK | | |
| | New Question (Subquestion) | What is the binary code of the first letter of the firstname of the author of Henry V in lowercase? | OK | | |
| | New Answer | 1110111 | | | |
| | Original Question | Who is the author of the play that was adapted into a film and featured the orchestral arrangement Suite from Henry V? | REFERENCE | | |
| | Original Answer | William Shakespeare | | | 5a81dacc55429926c1cdada0_2 |

Figure 6: Our annotation interface.

- question_decomposition: a list of sub-questions and sub-answers
- context: the two gold paragraphs
- answer_type: an answer type of the new question
- previous_answer_type: the answer type of the previous 2-hop question
- no_of_hops: the number of hops in our extended question
- reasoning_type: the list of required reasoning types
- pattern: a template that is used to generate the new question
- subquestion_patterns: a list of sub-questions of the template that is used to generate the new question
- cutted_question: the noun form that we obtain from the previous 2-hop question
- ques_on_last_hop: instead of integrating the new hop into the entire previous 2-hop question, we integrate it into the second hop of the previous 2-hop question. This is the third case (Case 3) in Figure 2.

We present an example in our dataset in Table 5.

**id:** 2hop__623931_656446
**question:** Who is the spouse of a cast member of Secrets of a Windmill Girl?
**answer:** John Alderton
**question_decomposition:**
- sub question 1: Secrets of a Windmill Girl » cast member
- sub answer 1: Pauline Collins
- sub paragraph_support_title 1: Secrets of a Windmill Girl
- sub question 2: #1 » spouse
- sub answer 2: John Alderton
- sub paragraph_support_title 2: Mrs Caldicot's Cabbage War

**context:**
*Paragraph 1: Secrets of a Windmill Girl*
Secrets of a Windmill Girl is a 1966 British exploitation film directed by Arnold L Miller. It recounts the road to ruin of a young woman (Pauline Collins) who becomes involved with the striptease scene after becoming a dancer at the Windmill Theatre in London. The film features fan dances by former Windmill Theatre Company performers. It was originally released in Britain as part of a double bill with N̈aked as Nature Intended̈.
*Paragraph 2: Mrs Caldicot's Cabbage War*
Mrs Caldicot's Cabbage War is a British comedy-drama film from 2002, directed by Ian Sharp and starring Pauline Collins, John Alderton and Peter Capaldi. It is based on a 1993 novel with the same name by Vernon Coleman.

Table 3: This is an example in MuSiQue where we do not have enough evidence to infer that the final answer (the spouse of Pauline Collins) is John Alderton.

### A.6 Dataset Analysis

As mentioned in Section 3, there are four answer types in our dataset: date, number, string, and letter. We present examples for each type of answer in Table 7.

Each sample in our dataset includes a list of question decompositions that can be useful for detailed analysis of the results. In addition, we include Case 3 (as shown in Figure 2), where we extend the second hop of the previous 2-hop question, rather than extending the entire previous 2-hop question. Currently, we use numbers to differentiate between these cases. The explanation for each case is as follows:

- Case 1: Our newly generated question
- Case 2: The previous 2-hop question
- Case 3: Our newly 2-hop generated question
- Case 4: Our extended question
- Case 5: The second hop of the previous 2-hop question
- Case 6: The first hop of the previous 2-hop question

In MoreHopQA w/ hv, we also ask humans to verify Case 3.

For 2Wiki and MuSiQue, the questions in Case 3 are automatically created using the same process as for questions in Case 1. In HotpotQA, to enhance efficiency, we use GPT-4 as the annotator to create the questions in Case 3.

## B Experiments

### B.1 Experimental Details

We run Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.3 and Gemma-7B-it on a single GPU (NVIDIA A100 40 GB), and Llama-3-70B-Instruct on 2 NVIDIA A100 80 GB GPUs. We use the following decoding parameters for all models: do_sample=True, max_new_tokens=256. The entire experiments took a total of 18 hours of runtime on the single GPU, and 30 hours on the pair of GPUs for LLama-3-70B. We additionally spent 84 $ to run GPT-4-Turbo. We wrote the Code for Evaluation with the help of Github Copilot.

**id:** 2hop__252311_366220
**question:** Who founded the company that distributed the film UHF?
**answer:** Mike Medavoy
**question_decomposition:**
- sub question 1: UHF » distributed by
- sub answer 1: Orion Pictures
- sub paragraph_support_title 1: UHF (film)
- sub question 2: #1 » founded by
- sub answer 2: Mike Medavoy
- sub paragraph_support_title 2: Mike Medavoy

**context:**
*Paragraph 1: UHF (film)*
Yankovic and Levey wrote the film after Yankovic's second studio album, looking to apply the musician's parody and comedy to film, and chose the approach of George being a straight man with a vivid imagination to support the inclusion of parodies within the film. They struggled with finding a film production company for financing the film, but were eventually able to get Orion Pictures' support after stating they could keep the film costs under $5 million. Principal filming took place around Tulsa, Oklahoma, with many of the extras for the film from the Tulsa and Dallas, Texas areas.
*Paragraph 2: Mike Medavoy*
Morris Mike Medavoy (born January 21, 1941) is an American film producer and executive, co-founder of Orion Pictures (1978), former chairman of TriStar Pictures, former head of production for United Artists (19742̌0131978) and current chairman and CEO of Phoenix Pictures.

Table 4: This is an example in MuSiQue. If we use the two provided paragraphs, the answer to the question is Mike Medavoy. However, if we do not use these paragraphs, there are multiple possible answers to the question because the Orion Pictures company was founded by five people: Arthur B. Krim, Eric Pleskow, Mike Medavoy, William Bernstein, and Robert Benjamin.

For NER in the postprocessing of the model answers as described in section 5.1, we used the NER module from spacy's en_core_web_sm pipeline. Please also see our published code for more details.

### B.2 Results

The full results are presented in Table 8.

### B.3 Performance Categorization

We present the details of the performance categorization in Table 9.

**_id:** fc0370920baf11ebab90acde48001122_14
**question:** What is the concatenation of the last letter of the first name and the first letter of the last name of the paternal grandmother of Mervyn Tuchet, 4Th Earl Of Castlehaven in lowercase?
**answer:** ym
**previous_question:** Who is the paternal grandmother of Mervyn Tuchet, 4Th Earl Of Castlehaven?
**previous_answer:** Lucy Mervyn

**question_decomposition:**
- sub question 1: Who is the father of Mervyn Tuchet, 4Th Earl Of Castlehaven?
- sub answer 1: Mervyn Tuchet, 2nd Earl of Castlehaven
- sub paragraph_support_title 1: Mervyn Tuchet, 4th Earl of Castlehaven
- sub question 2: Who is the mother of Mervyn Tuchet, 2Nd Earl Of Castlehaven?
- sub answer 2: Lucy Mervyn
- sub paragraph_support_title 2: Mervyn Tuchet, 2nd Earl of Castlehaven
- sub question 3: What is the concatenation of the last letter of the first name and the first letter of the lastname of Lucy Mervyn in lowercase?
- sub answer 3: ym
- sub paragraph_support_title 3:
- details: the details for the third sub-question

**context:**
*Paragraph 1: Mervyn Tuchet, 4th Earl of Castlehaven*
Mervyn Tuchet, 4th Earl of Castlehaven (died 2 November 1686) was the third son of Mervyn Tuchet, 2nd Earl of Castlehaven, and his first wife, Elizabeth Barnham (1592 – c. 1622)., He married Mary Talbot (buried 15 March 1710/1), daughter of John Talbot, 10th Earl of Shrewsbury (bef.,1601–1654) and his wife, née Mary Fortesque., ...
*Paragraph 2: Mervyn Tuchet, 2nd Earl of Castlehaven*
Mervyn Tuchet (sometimes Mervin Touchet), 2nd Earl of Castlehaven (1593 – 14 May 1631), was an English nobleman who was convicted of rape and sodomy and subsequently executed., A son of George Tuchet, 1st Earl of Castlehaven and 11th Baron Audley, by his wife, Lucy Mervyn, he was known by the courtesy title of Lord Audley during his father's lifetime, so is sometimes referred to as Mervyn Audley., ...
**answer_type:** string
**previous_answer_type:** person
**no_of_hops:** 5
**reasoning_type:** Symbolic, Commonsense
**pattern:** What is the concatenation of the last letter of the first name and the first letter of the last name of #Name in lowercase?
**subquestion_patterns:**
What is the first name of #Name?
What is the last letter of #Ans1?
What is the last name of #Name?
What is the first letter of #Ans3?
What is the concatenation of #Ans2 and #Ans4?
**cutted_question:** the paternal grandmother of Mervyn Tuchet, 4Th Earl Of Castlehaven
**ques_on_last_hop:** What is the concatenation of the last letter of the first name and the first letter of the lastname of the mother of Mervyn Tuchet, 2Nd Earl Of Castlehaven in lowercase?

Table 5: An example containing all information in our dataset. Due to the space limitation, we present the field 'details' in the 'question decomposition' part in Table 6.

| |
|---|
| **sub_id:** 3_1<br>**question:** What is the first name of Lucy Mervyn?<br>**answer:** Lucy |
| **sub_id:** 3_2<br>**question:** What is the last letter of Lucy?<br>**answer:** y |
| **sub_id:** 3_3<br>**question:** What is the last name of Lucy Mervyn?<br>**answer:** Mervyn |
| **sub_id:** 3_4<br>**question:** What is the first letter of Mervyn?<br>**answer:** m |
| **sub_id:** 3_5<br>**question:** What is the concatenation of y and m?<br>**answer:** ym |

Table 6: Example of the field 'details' in the 'question decomposition' part in Table 5.

| Question | Answer | Type |
|---|---|---|
| What is the date one day after when Prince Nikolai Of Denmark's mother was born? | 1964-07-01 | Date |
| How many letters are there between the first and last letters of the first name of the director of a 2004 film where Kam Heskin plays Paige Morgan in? | 4 | Number |
| What is the alphabetical order of the letters in the last name of the father of the director of film My 20Th Century? | deeiny | String |
| What is the last letter of the last name of the father of Empress Wang's husband? | i | Letter |

Table 7: Examples of different answer types in our dataset.

| Model | Case 1 | | Case 2 | | Case 3 | | Case 4 | | Case 5 | | Case 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 |
| Baseline_zero-cot | $1.88_{\pm 0.81}$ | $7.93_{\pm 1.28}$ | $7.42_{\pm 1.70}$ | $20.12_{\pm 1.90}$ | | | | | | | | |
| Llama-8B_zeroshot | $23.26_{\pm 2.50}$ | $28.62_{\pm 2.50}$ | $66.99_{\pm 2.86}$ | $79.82_{\pm 2.01}$ | $26.03_{\pm 2.50}$ | $30.38_{\pm 2.59}$ | $35.69_{\pm 3.04}$ | $37.62_{\pm 2.86}$ | $85.33_{\pm 2.15}$ | $91.68_{\pm 1.35}$ | $75.31_{\pm 2.50}$ | $87.36_{\pm 1.61}$ |
| Llama-8B_2-shot | $24.06_{\pm 2.42}$ | $28.54_{\pm 2.45}$ | $66.91_{\pm 2.68}$ | $77.79_{\pm 2.18}$ | $28.18_{\pm 2.50}$ | $32.14_{\pm 2.56}$ | $35.69_{\pm 2.68}$ | $37.52_{\pm 2.54}$ | $86.31_{\pm 2.06}$ | $92.53_{\pm 1.25}$ | $78.26_{\pm 2.33}$ | $88.51_{\pm 1.56}$ |
| Llama-8B_3-shot | $23.17_{\pm 2.42}$ | $27.74_{\pm 2.54}$ | $71.20_{\pm 2.77}$ | $80.45_{\pm 2.03}$ | $28.35_{\pm 2.86}$ | $32.22_{\pm 2.71}$ | $37.66_{\pm 2.95}$ | $39.24_{\pm 3.01}$ | $86.94_{\pm 1.97}$ | $92.66_{\pm 1.24}$ | $78.18_{\pm 2.42}$ | $88.56_{\pm 1.55}$ |
| Llama-8B_zero-cot | $20.84_{\pm 2.59}$ | $26.48_{\pm 2.56}$ | $65.03_{\pm 2.68}$ | $78.07_{\pm 1.96}$ | $24.87_{\pm 2.68}$ | $29.38_{\pm 2.67}$ | $34.08_{\pm 2.86}$ | $36.36_{\pm 2.90}$ | $83.99_{\pm 2.33}$ | $90.69_{\pm 1.47}$ | $74.42_{\pm 2.50}$ | $86.76_{\pm 1.60}$ |
| Llama-8B_2-shot-cot | $28.26_{\pm 2.77}$ | $32.28_{\pm 2.65}$ | $69.14_{\pm 2.68}$ | $79.72_{\pm 1.99}$ | $34.08_{\pm 2.95}$ | $37.69_{\pm 2.78}$ | $45.97_{\pm 3.04}$ | $47.85_{\pm 2.81}$ | $85.15_{\pm 2.06}$ | $91.80_{\pm 1.39}$ | $77.82_{\pm 2.50}$ | $88.04_{\pm 1.52}$ |
| Llama-8B_3-shot-cot | $30.50_{\pm 2.59}$ | $34.38_{\pm 2.70}$ | $73.26_{\pm 2.59}$ | $82.07_{\pm 1.95}$ | $34.44_{\pm 2.86}$ | $38.12_{\pm 2.74}$ | $45.71_{\pm 2.95}$ | $47.26_{\pm 2.90}$ | $86.31_{\pm 2.15}$ | $92.16_{\pm 1.34}$ | $76.48_{\pm 2.42}$ | $86.26_{\pm 1.91}$ |
| Mistral-7B_zeroshot | $14.49_{\pm 2.06}$ | $20.87_{\pm 2.15}$ | $64.04_{\pm 2.77}$ | $73.85_{\pm 2.33}$ | $18.96_{\pm 2.42}$ | $24.36_{\pm 2.31}$ | $27.73_{\pm 2.68}$ | $30.59_{\pm 2.59}$ | $77.28_{\pm 2.50}$ | $83.13_{\pm 1.98}$ | $65.21_{\pm 2.68}$ | $78.72_{\pm 1.97}$ |
| Mistral-7B_2-shot | $17.17_{\pm 2.42}$ | $23.52_{\pm 2.42}$ | $69.68_{\pm 2.95}$ | $78.17_{\pm 2.26}$ | $22.90_{\pm 2.59}$ | $28.09_{\pm 2.59}$ | $33.72_{\pm 2.86}$ | $35.96_{\pm 2.81}$ | $84.53_{\pm 2.15}$ | $89.80_{\pm 1.68}$ | $76.39_{\pm 2.59}$ | $86.32_{\pm 1.70}$ |
| Mistral-7B_3-shot | $16.73_{\pm 2.15}$ | $23.17_{\pm 2.37}$ | $70.57_{\pm 2.86}$ | $78.37_{\pm 2.32}$ | $23.52_{\pm 2.50}$ | $28.40_{\pm 2.52}$ | $32.74_{\pm 2.95}$ | $35.17_{\pm 2.85}$ | $84.35_{\pm 2.24}$ | $89.92_{\pm 1.72}$ | $76.65_{\pm 2.59}$ | $86.29_{\pm 1.67}$ |
| Mistral-7B_zero-cot | $18.16_{\pm 2.24}$ | $23.94_{\pm 2.37}$ | $55.64_{\pm 2.86}$ | $68.33_{\pm 2.16}$ | $20.04_{\pm 2.50}$ | $25.40_{\pm 2.38}$ | $30.59_{\pm 2.77}$ | $33.90_{\pm 2.64}$ | $66.82_{\pm 2.95}$ | $77.15_{\pm 2.18}$ | $50.18_{\pm 3.04}$ | $70.78_{\pm 2.18}$ |
| Mistral-7B_2-shot-cot | $17.80_{\pm 2.33}$ | $23.88_{\pm 2.46}$ | $68.96_{\pm 2.68}$ | $77.48_{\pm 2.14}$ | $24.87_{\pm 2.59}$ | $29.97_{\pm 2.61}$ | $37.48_{\pm 2.86}$ | $40.15_{\pm 2.91}$ | $85.51_{\pm 2.15}$ | $90.76_{\pm 1.55}$ | $75.85_{\pm 2.59}$ | $85.64_{\pm 1.98}$ |
| Mistral-7B_3-shot-cot | $19.41_{\pm 2.42}$ | $25.75_{\pm 2.44}$ | $68.34_{\pm 2.77}$ | $76.92_{\pm 2.19}$ | $25.94_{\pm 2.59}$ | $31.12_{\pm 2.67}$ | $37.57_{\pm 2.77}$ | $40.15_{\pm 2.84}$ | $85.69_{\pm 2.15}$ | $91.00_{\pm 1.52}$ | $75.49_{\pm 2.59}$ | $85.69_{\pm 1.80}$ |
| Gemma-7B_zeroshot | $7.07_{\pm 1.52}$ | $12.81_{\pm 1.78}$ | $40.07_{\pm 3.04}$ | $49.24_{\pm 2.62}$ | $9.48_{\pm 1.79}$ | $14.77_{\pm 1.82}$ | $18.87_{\pm 2.42}$ | $24.52_{\pm 2.32}$ | $59.12_{\pm 2.95}$ | $69.91_{\pm 2.35}$ | $52.86_{\pm 2.86}$ | $69.78_{\pm 2.15}$ |
| Gemma-7B_2-shot | $11.27_{\pm 1.88}$ | $16.85_{\pm 2.04}$ | $32.83_{\pm 2.86}$ | $41.09_{\pm 2.55}$ | $13.15_{\pm 1.97}$ | $18.11_{\pm 2.15}$ | $21.74_{\pm 2.59}$ | $26.31_{\pm 2.56}$ | $50.89_{\pm 3.04}$ | $61.64_{\pm 2.46}$ | $40.52_{\pm 2.95}$ | $57.63_{\pm 2.31}$ |
| Gemma-7B_3-shot | $8.94_{\pm 1.70}$ | $14.71_{\pm 1.81}$ | $27.91_{\pm 2.68}$ | $37.41_{\pm 2.55}$ | $12.52_{\pm 2.06}$ | $17.76_{\pm 2.04}$ | $22.09_{\pm 2.59}$ | $26.63_{\pm 2.68}$ | $44.99_{\pm 3.04}$ | $55.99_{\pm 2.62}$ | $34.70_{\pm 2.77}$ | $52.00_{\pm 2.27}$ |
| Gemma-7B_zero-cot | $7.33_{\pm 1.52}$ | $13.23_{\pm 1.73}$ | $33.81_{\pm 2.86}$ | $43.80_{\pm 2.62}$ | $10.02_{\pm 1.79}$ | $15.73_{\pm 1.86}$ | $15.74_{\pm 2.24}$ | $21.81_{\pm 2.26}$ | $49.73_{\pm 3.13}$ | $61.99_{\pm 2.58}$ | $39.53_{\pm 2.86}$ | $58.44_{\pm 2.29}$ |
| Gemma-7B_2-shot-cot | $10.55_{\pm 1.97}$ | $15.46_{\pm 1.84}$ | $31.57_{\pm 2.59}$ | $39.76_{\pm 2.52}$ | $13.51_{\pm 2.06}$ | $18.00_{\pm 2.18}$ | $10.91_{\pm 1.88}$ | $16.62_{\pm 1.97}$ | $59.48_{\pm 3.13}$ | $69.11_{\pm 2.63}$ | $42.84_{\pm 3.04}$ | $59.71_{\pm 2.34}$ |
| Gemma-7B_3-shot-cot | $10.82_{\pm 1.88}$ | $15.04_{\pm 1.91}$ | $24.15_{\pm 2.50}$ | $32.85_{\pm 2.44}$ | $11.90_{\pm 1.97}$ | $16.92_{\pm 2.16}$ | $11.99_{\pm 1.97}$ | $16.54_{\pm 2.06}$ | $56.53_{\pm 3.13}$ | $66.84_{\pm 2.44}$ | $40.34_{\pm 2.86}$ | $56.83_{\pm 2.43}$ |
| GPT-4_zeroshot | $51.33_{\pm 8.67}$ | $53.11_{\pm 8.33}$ | $87.33_{\pm 6.00}$ | $91.29_{\pm 4.22}$ | $64.00_{\pm 8.00}$ | $65.78_{\pm 7.78}$ | $67.33_{\pm 8.00}$ | $67.67_{\pm 7.67}$ | $90.00_{\pm 5.33}$ | $92.81_{\pm 4.30}$ | $70.67_{\pm 7.33}$ | $83.86_{\pm 4.93}$ |
| GPT-4_2-shot | $72.00_{\pm 7.33}$ | $73.44_{\pm 7.00}$ | $88.00_{\pm 6.00}$ | $90.91_{\pm 4.93}$ | $72.00_{\pm 7.33}$ | $74.11_{\pm 6.67}$ | $74.00_{\pm 7.33}$ | $74.33_{\pm 7.33}$ | $90.67_{\pm 5.33}$ | $92.65_{\pm 4.38}$ | $74.67_{\pm 7.33}$ | $86.08_{\pm 4.89}$ |
| GPT-4_3-shot | $68.67_{\pm 7.33}$ | $70.11_{\pm 7.44}$ | $88.67_{\pm 5.33}$ | $91.05_{\pm 4.49}$ | $69.33_{\pm 7.33}$ | $70.80_{\pm 7.20}$ | $80.00_{\pm 6.67}$ | $80.00_{\pm 6.67}$ | $86.67_{\pm 5.33}$ | $89.31_{\pm 4.49}$ | $76.00_{\pm 6.67}$ | $86.26_{\pm 4.97}$ |
| GPT-4_zero-cot | $72.67_{\pm 7.33}$ | $72.70_{\pm 7.30}$ | $88.00_{\pm 5.33}$ | $91.69_{\pm 4.33}$ | $74.67_{\pm 7.33}$ | $76.67_{\pm 6.67}$ | $81.33_{\pm 6.67}$ | $81.33_{\pm 6.67}$ | $90.00_{\pm 4.67}$ | $92.43_{\pm 3.91}$ | $67.33_{\pm 8.00}$ | $81.51_{\pm 5.18}$ |
| GPT-4_2-shot-cot | $73.33_{\pm 7.33}$ | $74.44_{\pm 7.11}$ | $88.67_{\pm 5.33}$ | $92.32_{\pm 3.90}$ | $77.33_{\pm 7.33}$ | $79.02_{\pm 6.62}$ | $82.00_{\pm 7.33}$ | $81.67_{\pm 7.00}$ | $90.00_{\pm 4.67}$ | $91.98_{\pm 4.27}$ | $72.67_{\pm 7.33}$ | $83.95_{\pm 4.88}$ |
| GPT-4_3-shot-cot | $72.00_{\pm 7.33}$ | $73.13_{\pm 6.87}$ | $90.67_{\pm 5.33}$ | $93.54_{\pm 4.16}$ | $75.33_{\pm 6.67}$ | $76.78_{\pm 6.69}$ | $78.00_{\pm 6.67}$ | $78.42_{\pm 6.60}$ | $90.67_{\pm 4.67}$ | $93.09_{\pm 3.93}$ | $72.00_{\pm 7.33}$ | $84.71_{\pm 5.01}$ |
| Llama-70B_zeroshot | $36.23_{\pm 3.04}$ | $38.46_{\pm 3.00}$ | $82.56_{\pm 2.33}$ | $90.18_{\pm 1.43}$ | $39.18_{\pm 3.04}$ | $41.06_{\pm 2.95}$ | $50.63_{\pm 3.13}$ | $50.79_{\pm 3.16}$ | $87.75_{\pm 1.97}$ | $92.96_{\pm 1.25}$ | $77.55_{\pm 2.50}$ | $88.03_{\pm 1.58}$ |
| Llama-70B_2-shot | $38.10_{\pm 2.86}$ | $39.80_{\pm 2.82}$ | $85.69_{\pm 2.06}$ | $92.23_{\pm 1.26}$ | $41.86_{\pm 2.95}$ | $43.43_{\pm 2.86}$ | $52.68_{\pm 3.13}$ | $52.77_{\pm 3.22}$ | $89.53_{\pm 1.88}$ | $93.89_{\pm 1.30}$ | $80.41_{\pm 2.42}$ | $89.59_{\pm 1.55}$ |
| Llama-70B_3-shot | $38.64_{\pm 2.95}$ | $40.30_{\pm 3.04}$ | $87.21_{\pm 1.97}$ | $92.96_{\pm 1.34}$ | $40.97_{\pm 2.86}$ | $42.40_{\pm 2.91}$ | $52.24_{\pm 3.31}$ | $52.59_{\pm 3.26}$ | $89.71_{\pm 1.88}$ | $94.14_{\pm 1.30}$ | $80.50_{\pm 2.33}$ | $89.76_{\pm 1.56}$ |
| Llama-70B_zero-cot | $49.91_{\pm 2.95}$ | $51.29_{\pm 3.06}$ | $80.95_{\pm 2.33}$ | $88.71_{\pm 1.57}$ | $51.79_{\pm 3.04}$ | $53.35_{\pm 3.09}$ | $56.17_{\pm 2.95}$ | $56.48_{\pm 3.00}$ | $88.01_{\pm 1.97}$ | $93.27_{\pm 1.23}$ | $77.37_{\pm 2.50}$ | $88.31_{\pm 1.61}$ |
| Llama-70B_2-shot-cot | $59.21_{\pm 2.86}$ | $60.06_{\pm 2.77}$ | $88.28_{\pm 1.97}$ | $94.33_{\pm 1.13}$ | $58.94_{\pm 3.04}$ | $60.20_{\pm 2.95}$ | $66.73_{\pm 2.77}$ | $66.86_{\pm 2.86}$ | $89.53_{\pm 1.79}$ | $93.97_{\pm 1.26}$ | $79.96_{\pm 2.50}$ | $89.38_{\pm 1.49}$ |
| Llama-70B_3-shot-cot | $57.51_{\pm 2.86}$ | $58.47_{\pm 2.91}$ | $87.57_{\pm 1.88}$ | $93.66_{\pm 1.16}$ | $59.12_{\pm 3.04}$ | $60.35_{\pm 3.03}$ | $66.01_{\pm 2.77}$ | $66.50_{\pm 2.77}$ | $89.62_{\pm 1.88}$ | $94.07_{\pm 1.20}$ | $80.59_{\pm 2.50}$ | $89.79_{\pm 1.51}$ |

Table 8: EM and F1 scores of the models on our dataset, together with $95\%$-confidence intervals obtained from bootstrapping ($n = 1000$) on the dataset. It is noted that the scores from GPT-4 are based on 150 samples (similar to the subset used for human performance), while for others, they are based on the full version of MoreHopQA w/ hv. The baseline model is Llama-8B prompted with the full context and only the first two words of the question.

<table>
<thead>
<tr><th>Case 1</th><th>Case 2</th><th>Case 3</th><th>Case 4</th><th>Case 5</th><th>Case 6</th><th>Category</th></tr>
</thead>
<tbody>
<tr><td rowspan="4">T</td><td rowspan="3">T</td><td>T</td><td>T</td><td>T</td><td>T</td><td>Perfect Reasoning</td></tr>
<tr><td colspan="2">-</td><td>F<br>T<br>F</td><td>F<br>F<br>T</td><td>Shortcut Reasoning</td></tr>
<tr><td colspan="2">Either is F</td><td>T</td><td>T</td><td>Problematic Performance</td></tr>
<tr><td>F</td><td colspan="4">-</td><td>Problematic Performance</td></tr>
<tr><td rowspan="9">F</td><td rowspan="4">T</td><td colspan="2">-</td><td>F<br>T<br>F</td><td>F<br>F<br>T</td><td>Shortcut Reasoning</td></tr>
<tr><td>T</td><td>F</td><td>T</td><td>T</td><td>Problematic Performance</td></tr>
<tr><td>-</td><td>T</td><td>T</td><td>T</td><td>Failed Reasoning</td></tr>
<tr><td>F</td><td>F</td><td>T</td><td>T</td><td>Extra Step Failure</td></tr>
<tr><td rowspan="5">F</td><td rowspan="3">T</td><td>F<br>T<br>F</td><td>F<br>F<br>T</td><td>-</td><td>Problematic Performance</td></tr>
<tr><td rowspan="2">T</td><td rowspan="2">T</td><td>T</td><td>Failed Reasoning</td></tr>
<tr><td>F</td><td>Failure</td></tr>
<tr><td rowspan="2">F</td><td>T</td><td>T</td><td>-</td><td>Failed Reasoning</td></tr>
<tr><td>F<br>T<br>F</td><td>F<br>F<br>T</td><td>-</td><td>Failure</td></tr>
</tbody>
</table>

Table 9: Categorizing the performance of the LLMs across various cases. T (true) means the LLM gives a correct answer to corresponding cases, while F (false) means the LLM gives a wrong one.