

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 DO LLMS PERFORM MULTILINGUAL MULTI-STEP REASONING?

Anonymous authors

Paper under double-blind review

ABSTRACT

Ideally, large language models (LLMs) should be able to exploit information sources from all available languages to achieve strong performance for diverse tasks, including reasoning. However, most evaluations of multilingual reasoning focus on symbolic domains, e.g., mathematics and coding, and it remains unclear how effective LLMs handle multilingual reasoning in linguistic tasks. In this paper, we introduce a controlled multilingual two-hop question answering setting, where answering a question requires two reasoning steps across two documents in different languages: the first-hop document provides bridging information, and the second-hop document links it to the final answer. Despite the equal importance of both hops, we find that the performance of a strong multilingual LLM (i.e., Gemma-3) is substantially affected by language variation in the second-hop documents more than in the first-hop. To analyze each hop’s reasoning process, we evaluate the decomposed sub-questions of a two-hop question. Surprisingly, the model often fails on the first sub-question for inferring bridging entities, yet still answers the overall two-hop question correctly. Our implicit context attribution analysis shows that the model still attends to bridging documents for correct answer generation, despite struggling to interpret them. This shows that the LLM’s multilingual multi-hop reasoning does not follow a faithful step-by-step decomposition for sub-question answering. We also find that the absence of reasoning decomposition leads to about 18% composition failures, where both sub-questions are answered correctly while failing to answer the two-hop question. To mitigate this, we propose a three-stage SUBQ prompting method to guide the multi-step reasoning with sub-questions, which boosts accuracy from 10.1% to 66.5%. Overall, our findings shed light on the multilingual multi-step reasoning mechanism and the potential of explicit reasoning decomposition for future tasks.

1 INTRODUCTION

Reasoning is a central aspect of human cognition and refers to the process of drawing new conclusions by combining multiple pieces of evidence through logical inference (Kurtz et al., 1999). Large language models (LLMs) have demonstrated strong, monolingual reasoning performance across pieces of evidence within the same language, which almost exclusively is English (Liu et al., 2025). However, real-world information is inherently multilingual and therefore distributed across languages. Enhancing multilingual reasoning is thus essential for building globally reliable AI systems, requiring models not only to comprehend information in multiple languages but also to integrate and chain knowledge across them to derive correct answers (Ghosh et al., 2025).

Recent works focus on improving multilingual reasoning in non-linguistic domains such as mathematics and coding (Qin et al., 2023; Huang et al., 2023; Chai et al., 2024; Wang et al., 2025). However, far less is understood about multilingual multi-step reasoning in linguistic domains, which requires analyzing, synthesizing, and drawing inferences solely from textual information. Multi-hop reasoning is a representative task that requires integrating information across multiple contexts to answer a complex question. While extensively studied in monolingual English settings (Yang et al., 2024; Biran et al., 2024; Yu et al., 2025), it remains unclear how well LLMs perform multi-hop reasoning across multilingual documents – an important capability for real-world applications where information is naturally distributed across different languages and cultural sources.

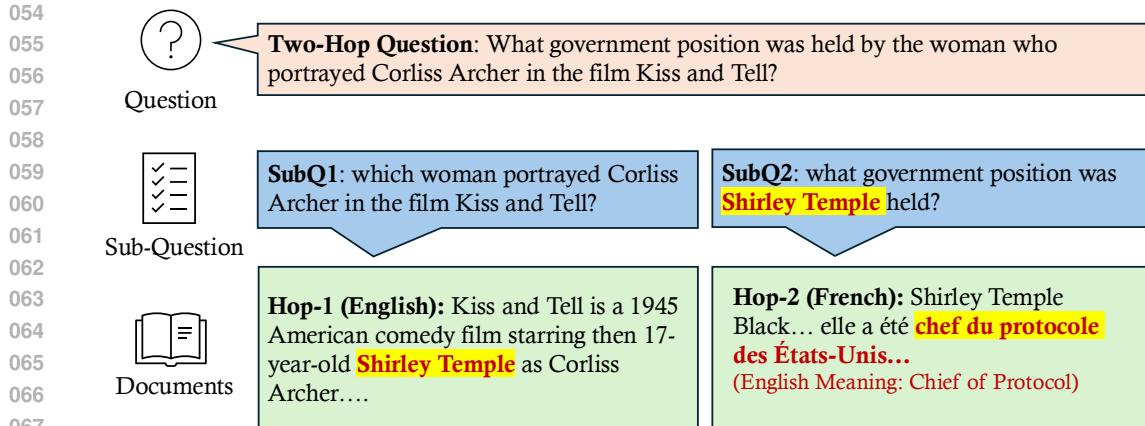


Figure 1: Example of *multilingual two-hop QA*. In Section 2, we evaluate multilingual two-step reasoning performance with the two-hop question and the corresponding Hop-1 and Hop-2 document. In Section 3, we conduct a sub-question evaluation to disentangle the two-step reasoning mechanism: SubQ1 infers the bridge entity and SubQ2 links the bridge entity to the final answer.

In this paper, we introduce a controlled multilingual two-hop question answering setting¹ to bridge the gap between the progress on English-based multi-hop reasoning and multilingual NLP. We extend a subset of the English-language HotpotQA dataset (Tang et al., 2021) by four diverse and high-resource languages: French, Russian, Arabic, and Chinese. Take Figure 1 as an example. To answer a two-hop question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?*, it requires two reasoning steps to infer the final answer: First, identifying the bridge entity *Shirley Temple* as the woman who portrayed the film from an English first-hop document. Second, connecting *Shirley Temple* with the government position *chief of protocol* in a French second-hop document. This setup provides a clear testbed to examine whether LLMs perform two-step multilingual reasoning and whether they reason in a step-by-step manner, commonly analogous to how humans solve problems via sub-questions.

We evaluate a strong multilingual large language model Gemma-3-Instruct 27B (Kamath et al., 2025) on this benchmark. From controlled experiments (in Section 2), we show a degradation in two-hop QA performance when varying languages for each document. In particular, changing languages in target-answer (second-hop) documents results in more performance drops compared to language changes in bridging (first-hop) documents. This performance discrepancy raises the question of whether the first step of multilingual reasoning is inherently easier than the second.

Subsequently, we conduct a step-level evaluation on decomposed sub-questions (in Section 3). We observe a relatively high unfaithfulness ratio in multilingual scenarios where the model correctly answers a two-hop question but fails to answer the first sub-question, with up to 33%. To further probe how the model arrives at the correct answer despite failing to identify bridging information, we conduct a context attribution analysis. The results reveal that the bridging documents still play a critical role in answer generation, with at least one-third of the attribution scores. The performance of unfaithful two-hop cases drops sharply when distractors with topically related bridging information are introduced. This shows that the LLM performs multilingual two-hop reasoning, but this reasoning does not necessarily indicate faithful step-by-step sub-question answering.

On the other hand, we show that faithful step-by-step sub-question answering does not necessarily ensure correct multilingual two-hop reasoning. Following Press et al. (2023), we refer to this phenomenon as compositional failure, and observe up to 18% of cases where the model correctly answers each sub-question but fails to integrate the intermediate information into the final two-hop answer. Next, we examine the impact of presenting explicit reasoning information on resolving this issue. We first show the existing chain-of-thought prompting (Wei et al., 2022), which requests models to “think step-by-step”, only partially alleviates compositional failure. With the given decomposed sub-questions, we naturally introduce a three-stage SUBQ prompting method, where each

¹We will release the dataset publicly to support reproducibility and further research.

Settings	Document Languages		Query Languages					Avg.
	Hop-1	Hop-2	En	Fr	Ru	Ar	Zh	
Monolingual	Q	Q	54.82	42.67	43.18	48.52	39.26	45.69
	\overline{Q}	Q	54.85	42.96	41.77	46.17	36.54	44.65
Multilingual	Q	\overline{Q}	43.07	36.88	39.73	42.12	34.78	39.31
	\overline{Q}	\overline{Q}	43.41	37.84	36.39	42.64	32.23	38.50

Table 1: Multilingual two-hop QA performance for Gemma-3-27B-Instruct. We report the F1 score token accuracy. Q/\overline{Q} denotes whether the languages of documents are the same or different from the query language. For languages of documents that differ from the query, we report their average performance. Full results are shown in Appendix A.5.

sub-question is explicitly provided to guide the final answer generation. This approach substantially reduces compositional failures, improving accuracy from 10.1% to 66.56%. These results highlight question decomposition as a promising direction for enhancing multilingual multi-step reasoning.

2 MULTILINGUAL TWO-HOP QUESTION ANSWERING

2.1 PRELIMINARY

The multilingual two-hop question answering (QA) task requires language models to reason with information from two gold documents in different languages to generate the answer. In this task, the models’ inputs are: (i) a two-hop question to answer and (ii) two gold documents. We denote Hop-1 as first-hop documents containing bridge entities and Hop-2 as second-hop documents with target answer information. In particular, a bridge entity must first be identified to infer the final answer in the second-hop documents. This task evaluates models’ ability on (a) multilingual understanding of gold documents, (b) multilingual reasoning ability to integrate information in varying languages.

2.2 EXPERIMENT SETUP

Dataset. We execute our task with the extended decomposed HotpotQA dataset (Tang et al., 2021), derived from (Yang et al., 2018). It comprises 1,000 English examples with two-hop questions and their corresponding decomposed sub-questions. Here, we use SubQ1 and SubQ2 to denote first and second single-hop sub-questions.

Dataset Filter. We apply a filtering procedure to the original dataset to mitigate the impact of data contamination. Specifically, we exclude data instances that can be correctly answered by the experimental models with partial or no gold documents: (a) only with Hop-1 or Hop-2 or (b) with neither Hop-1 nor Hop-2. This filtering ensures that the models rely on compositional reasoning over both hop documents rather than leveraging memorized knowledge from pre-training. The filtered multi-hop dataset contains a total of 182 examples.

Dataset Translation. We automatically translate the filtered English datasets into four high-resource languages with varying language families and written scripts: French (Fr), Russian (Ru), Arabic (Ar), and Chinese (Zh). We use GPT-4o-mini to translate the filtered English multi-hop datasets into four selected target languages. To ensure translation quality, we conducted human evaluations on a subset of translation examples (see Appendix A.2).

Models. We experiment with Gemma-3-Instruct (Kamath et al., 2025) model of size 27B due to its strong multilingual ability, which supports over 140 languages. We adopt greedy decoding to generate outputs and leave exploration of other decoding methods for future work. For the prompts, we put the two-hop question *before* and *after* the provided documents to reduce the effect of query-aware contextualization (Liu et al., 2023). The standard prompt templates are in Appendix A.4.1.

Correlation	En		Fr		Ru		Ar		Zh	
	Hop-1	Hop-2								
Pearson	0.01	0.89	0.06	0.66	0.20	0.78	0.20	0.70	0.12	0.69
Spearman	0.09	0.70	0.01	0.60	0.09	0.60	0.12	0.50	0.01	0.50

Table 2: Pearson and Spearman correlations between two-hop QA performance and the linguistic similarity. Overall, linguistic similarity between two-hop questions and Hop-2 documents has a strong correlation with the final performance.

2.3 RESULTS

Table 1 presents the results of the multilingual two-hop QA performance when querying in different languages. In line with previous studies (Chua et al., 2024; Qi et al., 2025), language models always perform better at answering questions in English than other query languages under both monolingual and multilingual evaluation. In particular, we show that the multilingual reasoning performance is sensitive to language changes between different hop documents. The two-step reasoning performance drops more when changing languages in answer-span Hop-2 documents (avg. -6.38%) than in bridging Hop-1 documents (avg. -1.04%).

To examine the effect of linguistic similarity on final performance, we compute both Pearson and Spearman correlations between two-hop QA accuracy and the linguistic distance between questions and documents.² As shown in Table 2, two-hop QA performance shows a strong negative correlation with the linguistic distance between questions and Hop-2 documents. In particular, larger linguistic gaps between the Hop-2 documents and the questions consistently lead to lower accuracy.

The performance gap when changing languages in the first or second-hop documents shows that the model might process the first and second reasoning steps differently. In principle, changing languages in both documents, i.e., bridging and answer-span, should have a similar effect on final performance since they are equally essential to infer the final answer. To probe the underlying reasoning mechanism, Section 3 presents a step-level evaluation that assesses LLM performance on each reasoning step via the decomposed sub-questions.

3 MULTILINGUAL MULTI-STEP REASONING DECOMPOSITION

In this section, we disentangle the two-hop reasoning process by two single steps through our translated set of decomposed sub-questions (Tang et al., 2021). This explicit step-wise evaluation facilitates a more fine-grained understanding of multilingual two-hop reasoning behavior.

3.1 SETUP

We denote SubQ1 as the first-step sub-questions that extract bridge entities from the Hop-1 documents, and SubQ2 as the second-step sub-questions that retrieve final answers from the Hop-2 documents; see also Figure 1. Based on the decomposed evaluation, we examine whether multilingual multi-hop reasoning follows a faithful step-by-step decomposition. We further identify two distinct failure modes following the definition in previous works, i.e., *unfaithfulness* (Lyu et al., 2023) and *compositional failure* (Press et al., 2023), illustrated in Appendix Figure 9. We analyze these modes to better disentangle LLMs’ multilingual multi-step reasoning.

- **Unfaithfulness:** Models correctly answered a two-hop question while failing to answer its sub-questions, i.e., SubQ1 or SubQ2. The unfaithfulness ratio is calculated as the percentage of total unfaithful cases over all correctly answered two-hop questions. This ratio reflects whether the model is faithful to step-by-step decomposed sub-question reasoning.
- **Compositional Failure:** Models incorrectly answered two-hop questions while succeeding in both SubQ1 and SubQ2. The composition failure ratio is calculated as the percentage of total composition failure cases over all incorrectly answered two-hop questions. This ratio reflects the limitation of LLMs’ compositional reasoning.

²Details of linguistic similarity calculation are provided in Appendix A.3.

Setting	Answer Correctness			Query Languages				
	Two-HopQ	SubQ1	SubQ2	En	Fr	Ru	Ar	Zh
Monolingual	✓	✗	✓	0.07	0.19	0.12	0.10	0.16
	✓	✓	✗	0.02	0.00	0.00	0.05	0.04
	✓	✗	✗	0.03	0.00	0.02	0.00	0.00
Multilingual	✓	✗	✓	0.12	0.23	0.15	0.25	0.33
	✓	✓	✗	0.03	0.03	0.03	0.05	0.04
	✓	✗	✗	0.01	0.00	0.03	0.03	0.02

(a) Unfaithfulness Ratio

Monolingual	✗	✓	✓	0.14	0.16	0.03	0.17	0.18
Multilingual	✗	✓	✓	0.11	0.10	0.05	0.10	0.08

(b) Compositional Failure Ratio

Table 3: The decomposed sub-questions evaluation. We report the average ratio for multilingual settings. The full results are shown in Appendix A.6.

3.2 UNFAITHFUL FAILURE

3.2.1 OVERALL RESULTS

Table 3a shows that the model is more likely to be unfaithful to the first-step sub-question for both monolingual and multilingual settings. Consistent with Tang et al. (2021), we show that explicitly identifying the bridge entity is not required for the model to answer two-hop questions correctly in monolingual English settings. Furthermore, we find that multilingual settings yield higher unfaithfulness ratios than monolingual ones. In particular, Arabic and Chinese two-hop questions yield notably high unfaithfulness rates of 25% and 33%, respectively. To further probe how the LLM is still able to generate correct answers in unfaithful cases, we conduct a context utilization analysis to examine the role of both hop documents, especially for bridging Hop-1 documents. Specifically, we only focus on unfaithful cases that fail at the first sub-question, as they constitute the majority.

3.2.2 CONTEXT ATTRIBUTION ANALYSIS

Here, we aim to implicitly analyze how language models use different hop documents to generate the answers. We follow input attribution methods to measure how much each document contributes to generating the answers (Lundberg & Lee, 2017; Covert et al., 2021).

Setup. We use the input-erasure attribution method (Li et al., 2016) to measure the contribution of each hop’s document on the correct answer generation. Input erasure quantifies the contribution of an input component (typically a token) by measuring the change in the model’s probability on the ground truth when that component is removed. In our experiments, we compute token-level attributions with respect to producing the correct answer. A document’s attribution is obtained by summing the contributions of its tokens, after which we calculate its percentage attribution relative to the combined attribution of the two documents, i.e., Hop-1 and Hop-2. Finally, we report the average percentage attribution for each document across the entire data. To ensure a fair comparison, we examine the context attribution scores in unfaithful cases on the first sub-question against fully faithful ones,³ to show shifts in context attribution.

Results. Figure 2 presents Hop-1 attribution scores for faithful and unfaithful cases in English two-hop questions. In general, we surprisingly notice that most unfaithful cases exhibit relatively higher Hop-1 attribution scores compared to faithful cases. Although the LLM fails to explicitly infer bridge entities from Hop-1 documents for unfaithful cases, Hop-1 documents still implicitly contribute to the correct two-hop target answer generation. The relatively high Hop-1 attribution

³Two-hop, SubQ1, and SubQ2 all correct

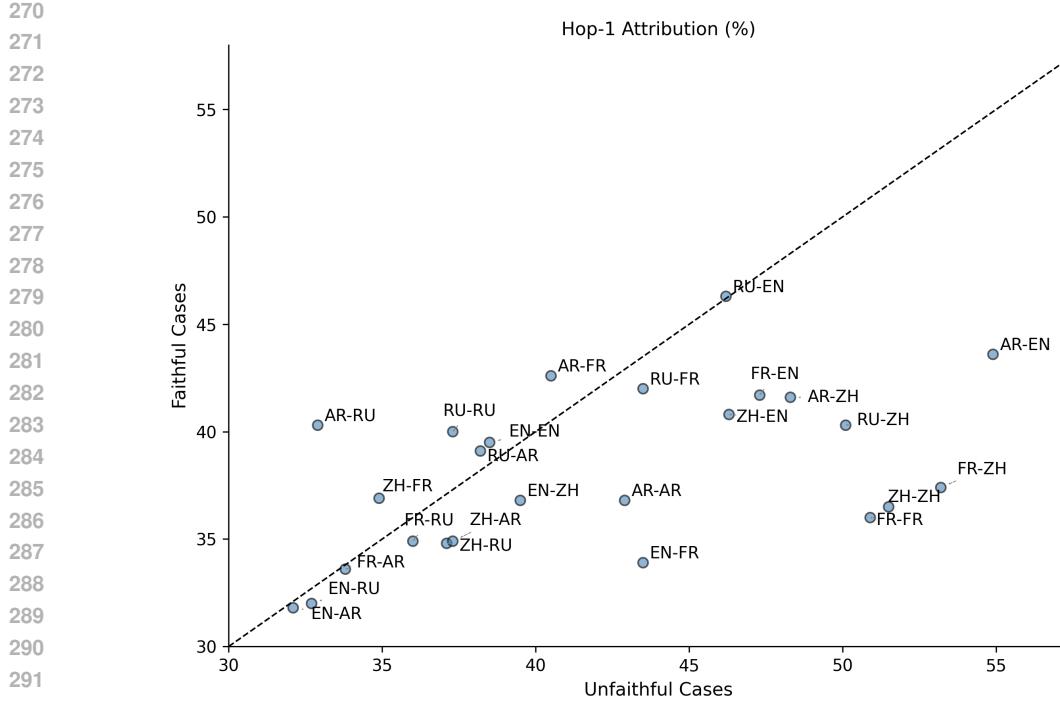


Figure 2: Context attribution scores for faithful and unfaithful cases. The two-hop query is in English, and Lang1-Lang2 (e.g., EN-ZH) indicates the languages of Hop-1 and Hop-2 documents.

scores for unfaithful cases indicate the model relies on a more nuanced intrinsic reasoning mechanism that extends beyond what can be revealed through explicit step-level evaluation. Moreover, we notice that multilingual Hop-1 documents have higher attributions than English ones for both cases. Taken together with the observation in Section 3.2.1 that multilingual settings exhibit higher unfaithfulness, this suggests that while the models attend to the multilingual Hop-1 documents, they struggle to answer the sub-question on them correctly.

3.2.3 CONTEXT DISTRACTOR ANALYSIS

The context attribution analysis shows the importance of Hop-1 documents on intrinsic multilingual reasoning for unfaithful cases. To further investigate whether the reliance on Hop-1 documents reflects robust reasoning rather than shortcircuiting spurious cues, we perform a controlled context perturbation analysis on unfaithful cases inspired by prior works by distractors (Hengle et al., 2025) and context orders (Yu et al., 2025).

Setup. We gradually insert an increasing number of distractor documents between Hop-1 and Hop-2 documents. The language of the distractors is the same as the two-hop question. We use distance d to denote the position differences between Hop-1 and Hop-2, and it corresponds to $(|d| - 1)$ distractors between the two hops. The sign of d specifies their order: a positive value means Hop-1 precedes Hop-2, while a negative value means the reverse. In particular, we control distractor relevance with the bridging Hop-1 documents. Relevant distractors contain topics similar to the original bridging Hop-1 documents from the dataset, whereas irrelevant distractors are randomly sampled from unrelated training examples.

Results. Figure 3 shows that inserting distractors between Hop-1 and Hop-2 degrades multilingual two-hop reasoning, consistent with monolingual findings by Modarressi et al. (2025). Topic-relevant distractors result in larger drops than irrelevant ones, underscoring the difficulty of the LLM to discriminate true bridging evidence from semantically proximate noise and highlighting the central role of Hop-1 in two-step reasoning.

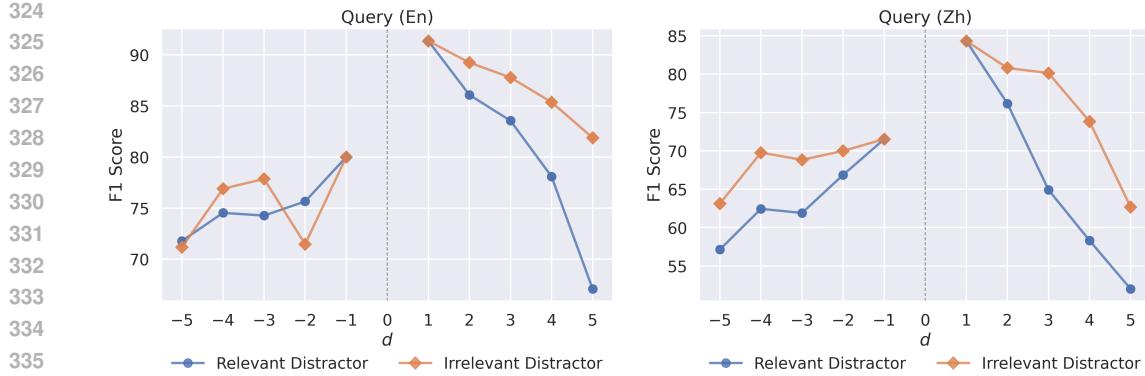


Figure 3: The impact of inserting relevant and irrelevant distractors between Hop-1 and Hop-2 documents. A distance of d corresponds to $(|d| - 1)$ distractors between the two hops. Positive d means Hop-1 precedes Hop-2, while a negative sign means the reverse. We report the average $F1$ token scores for every unfaithful multilingual case for each query language.

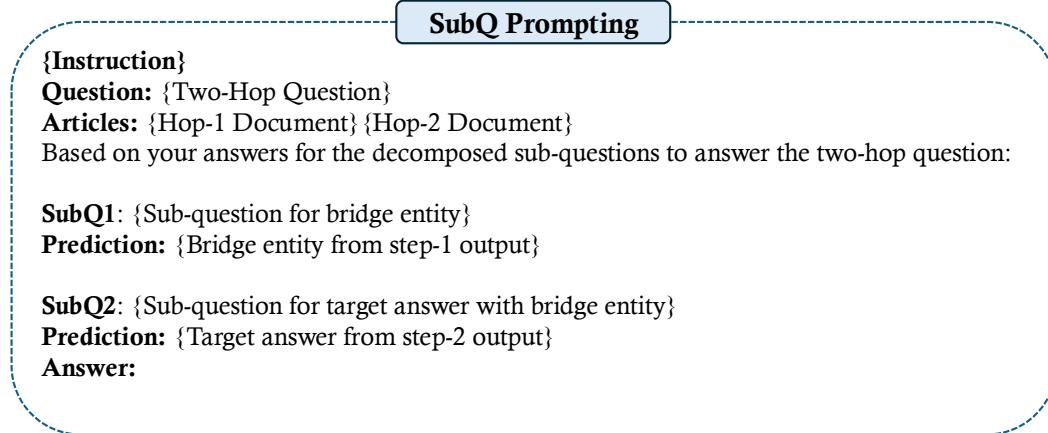


Figure 4: Three-step of SubQ Prompting. The first and second step prompt templates are shown in Appendix A.4.2 Figure 8.

Moreover, changing the document order further degrades performance: presenting the Hop-1 and Hop-2 documents in reverse order is consistently worse than the original sequence, echoing premise-order effects in LLMs (Chen et al., 2024; Yu et al., 2025). This indicates a positional sensitivity of intermediate bridge information, suggesting that the model relies on order-dependent reasoning rather than fully order-robust multi-hop composition.

3.3 COMPOSITIONAL FAILURE

The previous section 3.2 analyzes the unfaithfulness cases and shows that correct intrinsic reasoning does not necessarily follow faithful step-by-step sub-question answering. Here, we examine the second failure mode, i.e., compositional failure, where the model answers sub-questions correctly but fails to answer the overall two-hop question. Table 3b shows composition failure rates up to 18%, showing that solving both sub-questions does not guarantee correct two-hop composition. Motivated by evidence that prompting with explicit intermediate reasoning can improve performance (Wei et al., 2022), we aim to examine whether guiding the LLM with (i) self-generated chain-of-thoughts or (ii) decomposed step-by-step sub-questions, can reduce composition failure.

3.3.1 STEP-BY-STEP PROMPTING

The setup details and results are shown below.

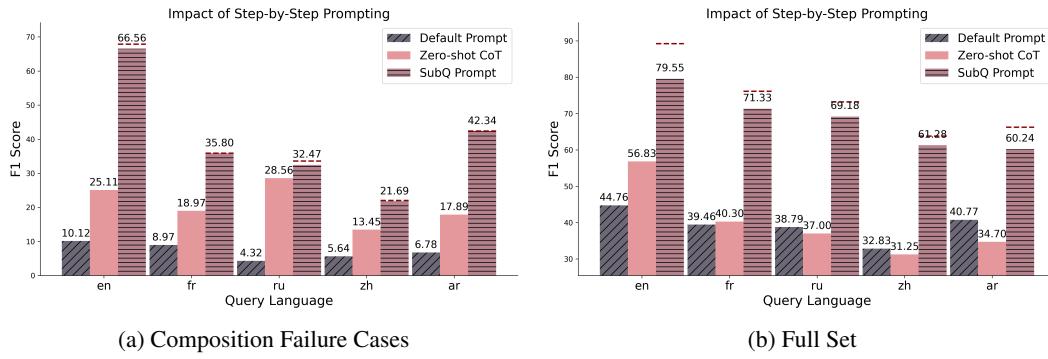


Figure 5: F1 token accuracy for multilingual two-hop QA for different prompting strategies, evaluated on composition failure cases and the full dataset. Red dashed lines above the SubQ Prompt bars show the performance when using ground-truth answers for sub-questions.

Zero-shot CoT. Following (Wei et al., 2022), we instruct the model to “Think Step-by-Step” and generate its own reasoning chains for answering the two-hop question (prompt templates are in Appendix A.4.2 Figure 7). The language of zero-shot instruction is aligned with the two-hop question, and we do not control the reasoning languages here.

SUBQ Prompt. We introduce a three-stage, step-by-step prompting technique with decomposed sub-questions. As illustrated in Figure 4, prompting consists of three steps: First step, prompting the model to answer the SubQ1 question about the bridge entity; second step, using the first step output for the bridge entity inserted with SubQ2 to ask for the target answer; third step, presenting both sub-questions and the previous steps’ outputs together to ask for the final two-hop answer. For a fair comparison, we also use the ground-truth answers for each step’s sub-question to guide reasoning.

Results. Figure 5 shows the average multilingual two-hop QA performance for each query language with different prompting techniques. First, SUBQ prompting yields substantial gains over zero-shot CoT on both the compositional failure cases (where both sub-questions are correctly answered but the final two-hop answer is wrong) and the full evaluation set. These improvements suggest that explicitly decomposing the query into sub-questions helps the model integrate information step-wise. One plausible explanation for the weaker zero-shot CoT results is that the model uses its own internal representations to generate explanations, which might diverge from the task’s true decomposition. This, in turn, leads to errors in the reasoning process. Future work could explore training or prompting with the task-specific reasoning decomposition.

Second, SUBQ prompting shows a larger performance gap between model-generated predictions and ground-truth answers for sub-questions on the full evaluation set compared to the compositional failure cases. This is expected, as the model predicts sub-questions correctly in compositional failure cases, resulting in performance close to that of the ground truth. For the full set, despite potential errors in predicted answers of sub-questions, SUBQ prompting still leads to substantial performance gains across all query languages. Overall, these results highlight the effectiveness of question decomposition in enhancing LLM performance on multilingual multi-hop reasoning tasks.

4 RELATED WORKS

In this section, we review prior works relevant to our study, focusing on multilingual capability, intrinsic reasoning, failure modes of reasoning, and the effectiveness of reasoning decomposition.

Multilingual Reasoning. This line of research has examined multilingual reasoning in mathematical tasks (Shi et al., 2023; Qin et al., 2023; Chen et al., 2023; Ghosh et al., 2025). Although mathematical problems also probe multi-step reasoning, these studies primarily focus on reasoning over numerical computation rather than general linguistic information. Another line of work has often noted linguistic benchmarks such as XCOPA (Ponti et al., 2020) or translated versions of

MMLU (Hendrycks et al., 2021) as multilingual reasoning (Chua et al., 2024). However, XCOPA typically involves one-step inferences, and MMLU lacks explicit control over reasoning steps. In contrast, our work introduces a controlled setting for multilingual multi-hop question answering, where we disentangle each step of reasoning across languages to directly assess whether models perform genuine multilingual linguistic reasoning.

Intrinsic Reasoning. Intrinsic reasoning investigates whether LLMs are capable of reasoning without explicit prompting (Wang & Zhou, 2024). For instance, Yang et al. (2024) reveals latent two-hop reasoning paths by constructing questions that require retrieving factual bridging information from pre-trained knowledge. Similarly, Guo et al. (2025) trains a three-layer transformer model from scratch to examine how two-hop reasoning emerges by analyzing attention logits for bridge and target entities. In contrast, our work evaluates reasoning explicitly through decomposed sub-questions, allowing us to assess whether multilingual multi-hop reasoning aligns with a step-by-step sub-question answering process.

Failure Modes of Reasoning. This work is also related to several modes of reasoning failure in LLMs, including unfaithfulness (Lyu et al., 2023; Arcuschin et al., 2025), premise order effects (Chen et al., 2024; Yu et al., 2025), distractibility (Guo et al., 2025; Qi et al., 2025), and the limited capacity for long-context reasoning (Liu et al., 2023; Modarressi et al., 2025; Hengle et al., 2025). Specifically, Lyu et al. (2023) shows that chain-of-thought prompting on mathematical tasks often produces answers that do not faithfully follow the intermediate reasoning steps. Our work extends to multi-hop reasoning by showing the unfaithful outputs for decomposed sub-questions. Modarressi et al. (2025) reveals that increasing the context length of the two-hop associated reasoning QA task leads to performance degradation, while we show that this difficulty is further amplified when long-context reasoning is combined with multilinguality, exposing another limitation of current LLMs.

Effectiveness of Reasoning Decomposition. Reasoning decomposition has been widely adopted in prompting techniques across various downstream tasks. For example, chain-of-thought prompting decomposes mathematical problems into intermediate steps, encouraging models to follow a step-by-step reasoning strategy (Wei et al., 2022). In machine translation, Briakou et al. (2024); He et al. (2024) decompose the task into multiple stages, showing that multi-turn refinements can improve translation quality. In our work, we explore multi-hop question decomposition. While the decomposed sub-questions need to be additionally acquired, they provide a potential way of enhancing multilingual multi-hop reasoning performance.

5 CONCLUSION

In this paper, we introduce a controlled setting for multilingual two-hop reasoning to broaden multilingual evaluation. Building on this task, we present a comprehensive analysis that exposes the limitations of a current strong multilingual LLM, i.e., Gemma-3, in two-step reasoning when pieces of evidence are cross-lingual. Specifically, we find that Gemma-3 is more sensitive to language variation in answer-span documents than in bridging documents, despite the equal importance of both for final answer generation. By decomposing each query into sub-questions, we disentangle the two-step reasoning process and observe that the model frequently fails the first sub-question (identifying the bridge entity) while still answering the overall two-hop question correctly. To probe unfaithfulness, we conduct context attribution analysis, which reveals that bridging information still plays a crucial role in final answer generation. Our findings indicate that multilingual multi-step reasoning in the LLM cannot be fully captured by human-like step-by-step sub-question answering, underscoring the need for future work to uncover the underlying mechanisms of model reasoning.

On the other hand, this work demonstrates the benefits of explicit step-by-step prompting with structured sub-questions in multilingual multi-hop tasks, particularly for mitigating compositional failures. Although the model is capable of reasoning without predefined structures, our findings show that such guidance enhances its robustness. These insights open promising directions for future work on incorporating reasoning decomposition into both training objectives and prompting strategies to improve multilingual multi-hop reasoning.

486 REFERENCES
487

488 Iv'an Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda,
489 and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *ArXiv*,
490 abs/2503.08679, 2025. URL <https://api.semanticscholar.org/CorpusID:276929360>.

491

492 Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late:
493 Exploring the limitations of large language models on multi-hop queries. In Yaser Al-Onaizan,
494 Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical
495 Methods in Natural Language Processing*, pp. 14113–14130, Miami, Florida, USA, November
496 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.781. URL
497 <https://aclanthology.org/2024.emnlp-main.781>.

498

499 Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. Translating step-by-step:
500 Decomposing the translation process for improved translation quality of long-form texts. In
501 Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the
502 Ninth Conference on Machine Translation*, pp. 1301–1317, Miami, Florida, USA, November
503 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.123. URL
504 <https://aclanthology.org/2024.wmt-1.123>.

505 Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang,
506 Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. xcot: Cross-lingual instruction tuning for
507 cross-lingual chain-of-thought reasoning. In *AAAI Conference on Artificial Intelligence*, 2024.
508 URL <https://api.semanticscholar.org/CorpusID:266999425>.

509

510 Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang,
511 and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and
512 observations, 2023.

513 Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in rea-
514 soning with large language models. *ArXiv*, abs/2402.08939, 2024. URL <https://api.semanticscholar.org/CorpusID:267657940>.

515

516 Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi,
517 Amer Sinha, Chulin Xie, and Chiyuan Zhang. Crosslingual capabilities and knowledge bar-
518 riers in multilingual large language models. *ArXiv*, abs/2406.16135, 2024. URL <https://api.semanticscholar.org/CorpusID:270703607>.

519

520 Ian Covert, Scott M. Lundberg, and Su-In Lee. Explaining by removing: A unified framework for
521 model explanation. *J. Mach. Learn. Res.*, 22:209:1–209:90, 2021. URL <https://jmlr.org/papers/v22/20-1316.html>.

522

523 Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. The multilingual mind : A
524 survey of multilingual reasoning in language models. *ArXiv*, abs/2502.09457, 2025. URL
525 <https://api.semanticscholar.org/CorpusID:276317681>.

526

527 Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I. Jordan, and Stuart Rus-
528 sell. How do llms perform two-hop reasoning in context? *ArXiv*, abs/2502.13913, 2025. URL
529 <https://api.semanticscholar.org/CorpusID:276449562>.

530

531 Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu,
532 Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language
533 models. *Transactions of the Association for Computational Linguistics*, 12:229–246, 2024. doi:
534 10.1162/tacl_a_00642. URL <https://aclanthology.org/2024.tacl-1.13>.

535

536 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Ja-
537 cob Steinhardt. Measuring massive multitask language understanding. In *International Confer-
538 ence on Learning Representations*, 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.

540 Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. Multilingual needle in a
 541 haystack: Investigating long-context behavior of multilingual large language models. In Luis
 542 Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations
 543 of the Americas Chapter of the Association for Computational Linguistics: Human Language
 544 Technologies (Volume 1: Long Papers)*, pp. 5165–5180, Albuquerque, New Mexico, April 2025.
 545 Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.
 546 naacl-long.267. URL <https://aclanthology.org/2025.naacl-long.267/>.

547

548 Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei.
 549 Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-
 550 thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the As-
 551 sociation for Computational Linguistics: EMNLP 2023*, pp. 12365–12394, Singapore, December
 552 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.826.
 553 URL <https://aclanthology.org/2023.findings-emnlp.826/>.

554 Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Mer-
 555 hej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouil-
 556 lard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec,
 557 Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, Francesco Visin, Kathleen Kenealy, Lucas
 558 Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Istvan Busa-Fekete, Alex Feng, Noveen Sachdeva,
 559 Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal,
 560 Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal,
 561 Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, An-
 562 drreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian
 563 Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit
 564 Vadi, Andr'as Gyorgy, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine
 565 Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Boxi Wu, Bobak Shahri-
 566 ari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey,
 567 Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas,
 568 Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Er-
 569 win Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi
 570 Hashemi, Hanna Klimczak-Pluci'nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hus-
 571 sein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan,
 572 Joe Stanton, J. Michael Weting, Jonathan Lai, Jordi Orbay, Joe Fernandez, Joshua Newlan,
 573 Junsong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff,
 574 Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Wat-
 575 son, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan
 576 Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron,
 577 Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Pi-
 578 otr Stańczyk, Pouya Dehghani Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Ardeshir
 579 Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Gir-
 580 gin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram
 581 Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra,
 582 Utku Evcı, Vedant Misra, Vincent Roseberry, Vladimir Feinberg, Vlad Kolesnikov, Woohyun
 583 Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor
 584 Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Mart-
 585 ins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab S. Mirrokni, Evan
 586 Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley,
 587 Slav Petrov, Noah Fiedel, Noam M. Shazeer, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray
 588 Kayukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry
 589 Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin,
 590 Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. *ArXiv*, abs/2503.19786,
 591 2025. URL <https://api.semanticscholar.org/CorpusID:277313563>.

592

593 Kenneth J. Kurtz, Dedre Gentner, and Virginia Gunn. Chapter 4 - reasoning1. In Ben-
 594jamin Martin Bly and David E. Rumelhart (eds.), *Cognitive Science*, pp. 145–200. Aca-
 595demic Press, San Diego, 1999. ISBN 978-0-12-601730-4. doi: <https://doi.org/10.1016/B978-012601730-4/50006-8>. URL <https://www.sciencedirect.com/science/article/pii/B9780126017304500068>.

594 Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation
595 erasure. *CoRR*, abs/1612.08220, 2016. URL <http://arxiv.org/abs/1612.08220>.

596

597 Hammeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue
598 Zhang. Logical reasoning in large language models: A survey, 2025. URL <https://arxiv.org/abs/2502.09100>.

599

600 Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni,
601 and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions
602 of the Association for Computational Linguistics*, 12:157–173, 2023. URL <https://api.semanticscholar.org/CorpusID:259360665>.

603

604 Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Is-
605 abelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vish-
606 wanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30:
607 Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long
608 Beach, CA, USA*, pp. 4765–4774, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.

609

610

611 Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki,
612 and Chris Callison-Burch. Faithful chain-of-thought reasoning. In Jong C. Park, Yuki Arase, Bao-
613 tian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of
614 the 13th International Joint Conference on Natural Language Processing and the 3rd Conference
615 of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long
616 Papers)*, pp. 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguis-
617 tics. doi: 10.18653/v1/2023.ijcnlp-main.20. URL <https://aclanthology.org/2023.ijcnlp-main.20>.

618

619 Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun
620 Yoon, and Hinrich Schütze. Nolima: Long-context evaluation beyond literal matching. In *Forty-
621 second International Conference on Machine Learning*, 2025. URL <https://arxiv.org/abs/2502.05167>.

622

623 Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen.
624 XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the
625 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
626 URL <https://ducdauge.github.io/files/xcopa.pdf>.

627

628 Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Mea-
629 suring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan
630 Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP
631 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics.
632 doi: 10.18653/v1/2023.findings-emnlp.378. URL <https://aclanthology.org/2023.findings-emnlp.378>.

633

634 Jirui Qi, Raquel Fernández, and Arianna Bisazza. On the consistency of multilingual context
635 utilization in retrieval-augmented generation. *ArXiv*, abs/2504.00597, 2025. URL <https://api.semanticscholar.org/CorpusID:277468130>.

636

637 Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompt-
638 ing: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor,
639 Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Meth-
640 ods in Natural Language Processing*, pp. 2695–2709, Singapore, December 2023. Associa-
641 tion for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.163. URL <https://aclanthology.org/2023.emnlp-main.163>.

642

643 Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework
644 for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Pro-
645 ceedings of the 2020 Conference on Empirical Methods in Natural Language Processing
646 (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguis-
647 tics. doi: 10.18653/v1/2020.emnlp-main.213. URL <https://aclanthology.org/2020.emnlp-main.213>.

648 Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi,
 649 Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Lan-
 650 guage models are multilingual chain-of-thought reasoners. In *The Eleventh International Confer-
 651 ence on Learning Representations*, 2023. URL <https://openreview.net/forum?id=fR3wGCk-IXp>.

652

653 Yixuan Tang, Hwee Tou Ng, and Anthony Tung. Do multi-hop question answering systems know
 654 how to answer the single-hop sub-questions? In Paola Merlo, Jorg Tiedemann, and Reut
 655 Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Associa-
 656 tion for Computational Linguistics: Main Volume*, pp. 3244–3249, Online, April 2021. Asso-
 657 ciation for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.283. URL <https://aclanthology.org/2021.eacl-main.283>.

658

659 Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. Multilingual prompting for im-
 660 proving llm generation diversity. *ArXiv*, abs/2505.15229, 2025. URL <https://api.semanticscholar.org/CorpusID:278782239>.

661

662 Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. *ArXiv*,
 663 abs/2402.10200, 2024. URL <https://api.semanticscholar.org/CorpusID:267681847>.

664

665 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le,
 666 and Denny Zhou. Chain of thought prompting elicits reasoning in large language mod-
 667 els. *ArXiv*, abs/2201.11903, 2022. URL <https://api.semanticscholar.org/CorpusID:246411621>.

668

669 Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large lan-
 670 guage models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and
 671 Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Com-
 672 putational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, Bangkok, Thailand, August
 673 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL
 674 <https://aclanthology.org/2024.acl-long.550>.

675

676 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov,
 677 and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question
 678 answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (eds.), *Proceed-
 679 ings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–
 680 2380, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
 681 doi: 10.18653/v1/D18-1259. URL <https://aclanthology.org/D18-1259>.

682

683 Sangwon Yu, Ik-hwan Kim, Jongyoon Song, Saehyung Lee, Junsung Park, and Sungroh Yoon.
 684 Unleashing multi-hop reasoning potential in large language models through repetition of mis-
 685 ordered context. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the As-
 686 sociation for Computational Linguistics: NAACL 2025*, pp. 6435–6455, Albuquerque, New
 687 Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7.
 688 doi: 10.18653/v1/2025.findings-naacl.360. URL <https://aclanthology.org/2025.findings-naacl.360>.

689

690

691

692

693

694

695

696

697

698

699

700

701

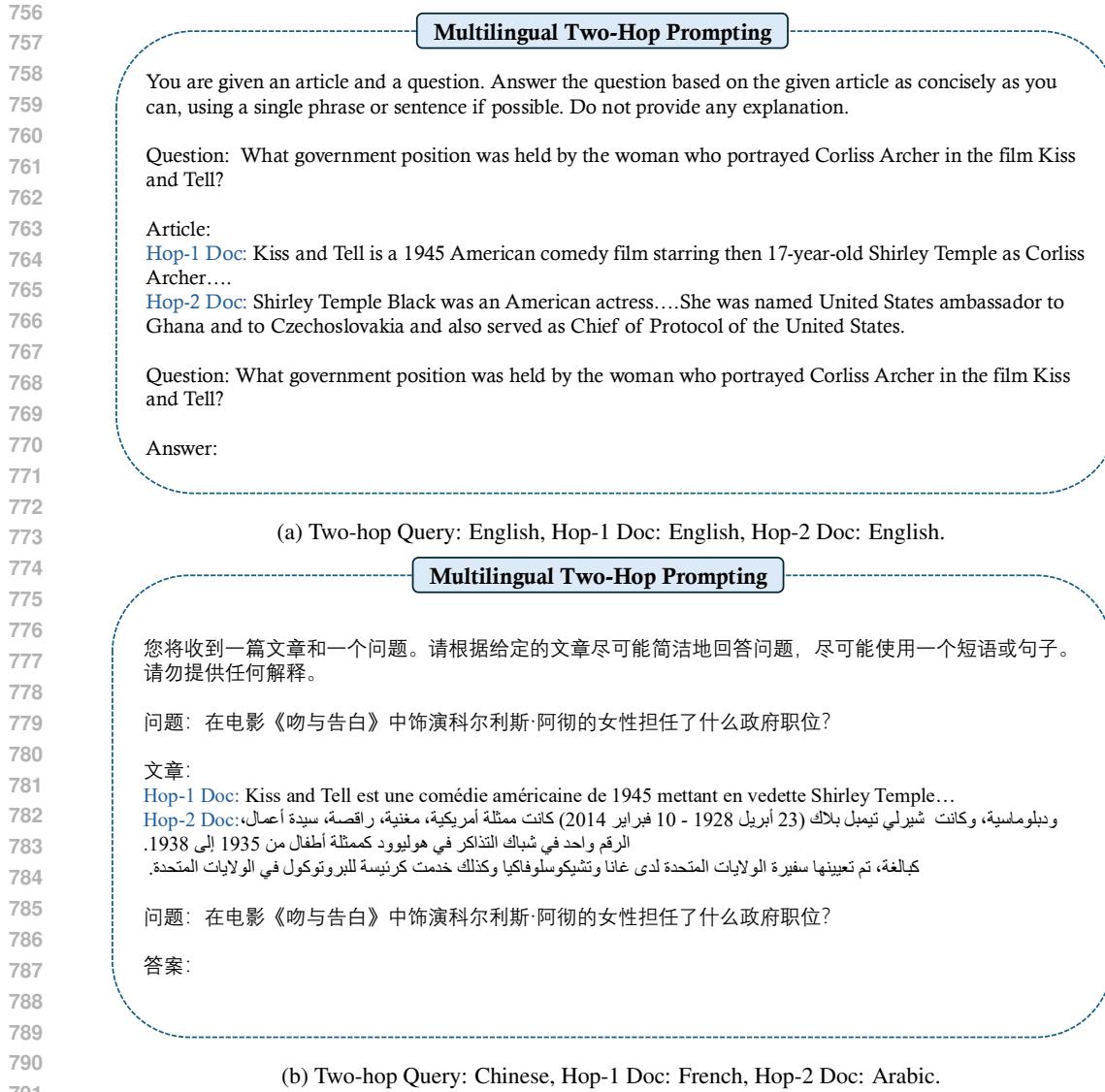
702
703 A APPENDIX704
705 A.1 THE USE OF LARGE LANGUAGE MODELS (LLMs)706
707 In this work, we use LLMs in two ways: First, we use GPT-4o-mini for dataset translation.
708 Second, we use ChatGPT solely for polishing the paper text, only for grammar corrections. We do
709 not use LLMs for research ideation or for generating substantive content for the paper.710
711 A.2 TRANSLATION QUALITY EVALUATION712
713 We evaluate the quality of the translation dataset using both automatic metrics and human judgments.
714 For automatic evaluation, we report reference-free COMET-22⁴ (Rei et al., 2020) scores. For human
715 evaluation, we randomly sample 20% of the dataset for each language and ask native speakers of
716 French, Chinese, Arabic, and Russian to rate the translations. Table 4 presents the human evaluation
717 results, and the corresponding rating criteria are described below:718
719

- 3: The translation conveys the same meaning as the source English, without grammar
errors.
- 2: The translation conveys most of the meanings with the source English. It contains a few
grammatical errors.
- 1: The translation only conveys some of the meanings in the English source. It might not
be fluent and may contain several grammatical errors.
- 0: The translation conveys little or no meaning in the source English, and the translation is
hard to understand.

720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

Metrics	French (Fr)	Russian (Ru)	Chinese (Zh)	Arabic (Ar)
COMET	86.14	83.42	82.89	80.43
Human	2.6	2.4	2.5	2.6

732
733 Table 4: Translation quality for the Multilingual Two-hop QA dataset. Overall, both COMET-based
734 automatic evaluation and human evaluation confirm that the translations are meaningful.735
736 A.3 LINGUISTIC SIMILARITY CALCULATION737
738 We calculate the linguistic similarity based on the subword vocabulary overlaps from a multi-parallel
739 corpus NTREX⁵ which covers 128 languages, followed by (Qi et al., 2025). We measure the sub-
740 word vocabulary overlap between language l_1 and l_2 as follows: $\text{Overlap}(l_1, l_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}$, where
741 V_1 and V_2 represent the subword vocabulary for language l_1 and l_2 .742
743 A.4 PROMPT TEMPLATES744
745 A.4.1 MULTILINGUAL TWO-HOP DEFAULT PROMPTING746
747 Figure 6 illustrates the default prompt templates used for both the monolingual and multilingual
748 two-hop tasks. For all the prompting, the language of the instruction is the same as the two-hop
749 question. We ask native speakers of each language, who are also proficient in English, to provide
750 the translations of the instructions.751
752 A.4.2 STEP-BY-STEP PROMPTING753
754 Figure 7 and 8 shows the step-by-step prompting techniques used in Section 3.3.1.755
756 ⁴Unbabel/wmt22-cometkiwi-da⁵<https://github.com/MicrosoftTranslator/NTREX>



810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

Zero-shot CoT

You are given an article and a question. Think step-by-step and then give a final concise answer.

Question: {TwoHopQ}
Article: {Hop-1} {Hop-2}
Question: {TwoHopQ}

Reasoning:

Answer:

Zero-shot CoT

你将获得一篇文章和一个问题。请逐步思考，然后给出最终的简洁答案。

问题: {TwoHopQ}
文章: {Hop-1} {Hop-2}
问题: {TwoHopQ}

推理:

答案:

(a) Two-hop Question in English.

(b) Two-hop Question in Chinese.

Figure 7: Zero-shot CoT prompt template. We add “think step-by-step” to prompt the models to generate their own reasoning chains.

Step 1 CoT

You are given an article and a question. Answer the question based on the given article as concisely as you can, using a single phrase or sentence if possible. Do not provide any explanation.

Question: {Sub-question for bridge entity}
Answer:

Step 2 CoT

You are given an article and a question. Answer the question based on the given article as concisely as you can, using a single phrase or sentence if possible. Do not provide any explanation.

Question: {Sub-question for target answer with Step 1’s predicted bridge entity}
Answer:

Figure 8: The first two steps of the three-stage SUBQ prompting. First step, answering the first sub-question to identify the bridge entity. Second step, using step 1’s predicted bridge entity to answer the second sub-question, and the final step, shown in the main text 4, combining both steps to generate the final two-hop answer.

Two-hopQ: Lou Cutell 出现在 CBS 情景喜剧第 九季 第三集，这一集在总体上是第几集?

Model Prediction: 187 (**correct**)

Ground-Truth: 187

Two-hopQ: Katherine Ann Dettwyler wasn’t rehired after she commented on the death of the student who was arrested in North Korea in what month and year?

Model Prediction: June 2017 (**wrong**)

Ground-Truth: January 2016

SubQ1: Lou Cutell 出现在哪一集?

Model Prediction: 螺旋形杰瑞 (**wrong**)

Ground-Truth: 最后一次在纽约

SubQ2: 最后一次在纽约是总体的第几集?

Model Prediction: 187 (**correct**)

Ground-Truth: 187

SubQ1: Katherine Ann Dettwyler wasn’t rehired after she commented on the death of which student?

Model Prediction: Otto Warmbier (**correct**)

Ground-Truth: Otto Warmbier

SubQ2: Otto Warmbier was arrested in North Korea in what month and year?

Model Prediction: January 2016 (**correct**)

Ground-Truth: January 2016

Figure 9: Two distinct reasoning failure modes from Gemma-3-27B-Instruct. **Left:** Unfaithfulness, **Right:** Composition Failure.

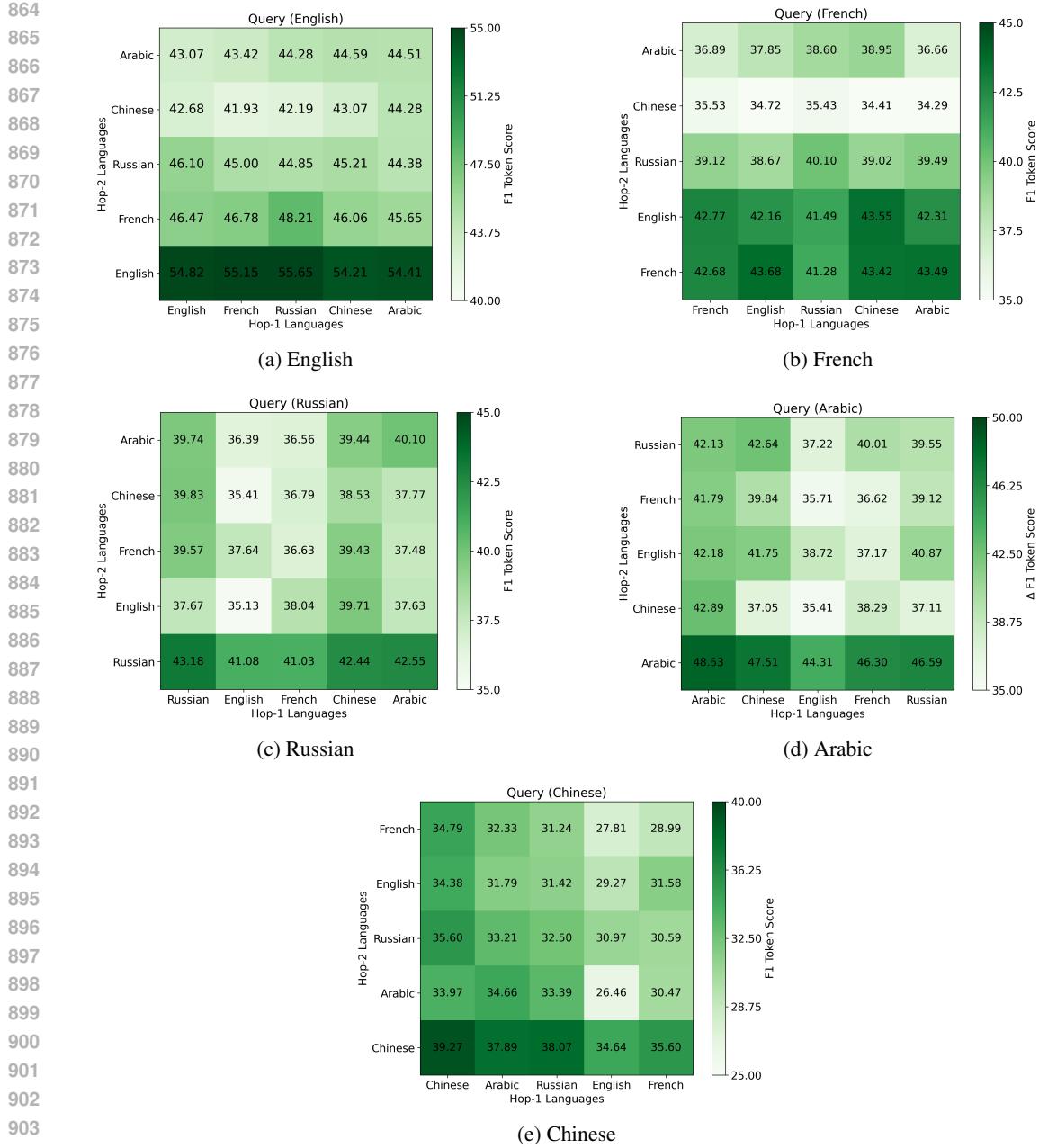


Figure 10: Multilingual Multi-Hop QA Performance for each query language with language variations in Hop-1 or Hop-2 documents for Gemma-3-27B-Instruct.

A.7 ADDITIONAL MODELS

Table 5 reports results for the Qwen-3-8B model, which are largely consistent with Gemma-3-27B-Instruct. Overall, the performance of LLMs is sensitive to language variation in the supporting documents: changing the language of the Hop-2 document causes a larger performance drop than changing the language of the Hop-1 document.

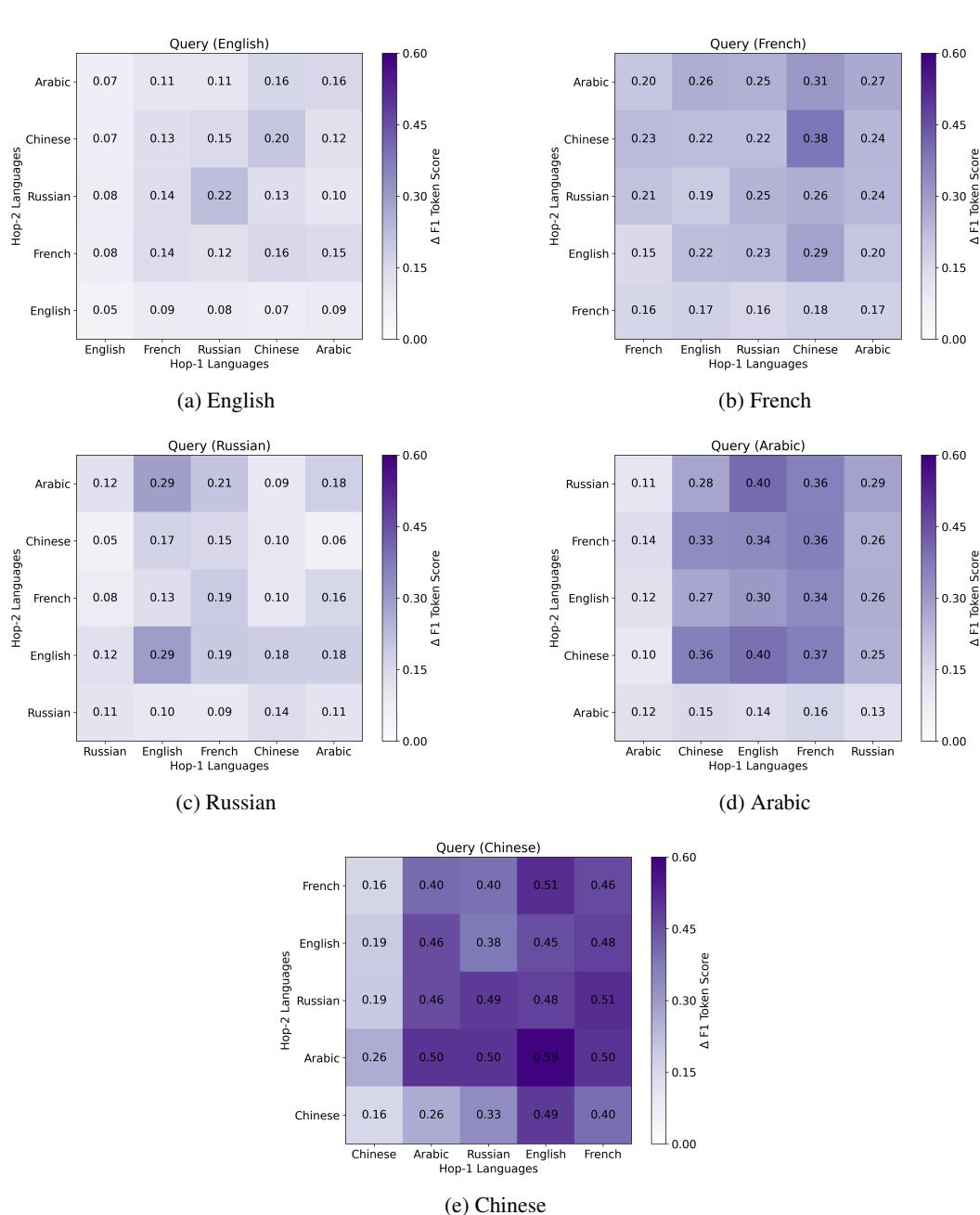


Figure 11: Unfaithfulness Ratios

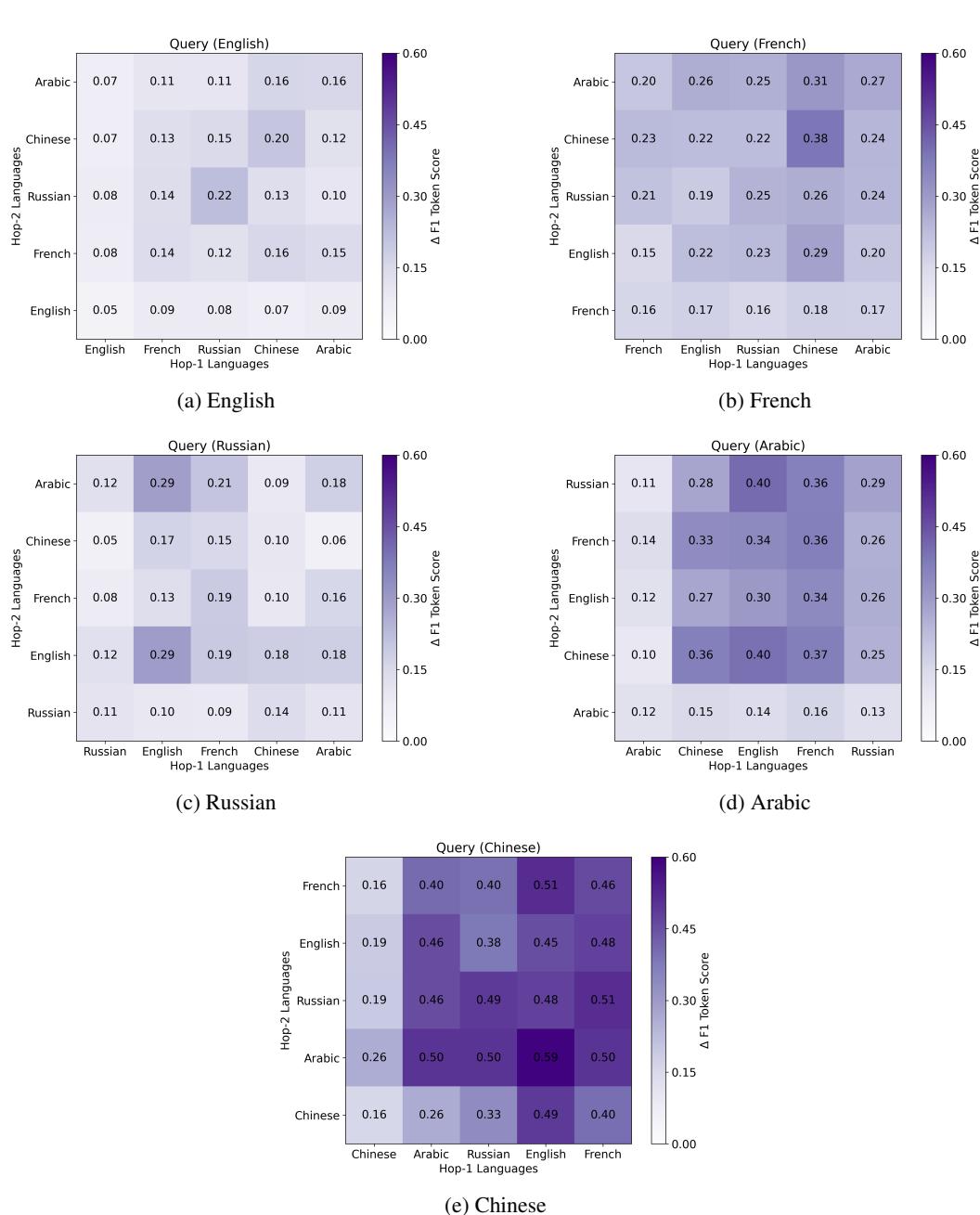


Figure 12: Composition Failure Ratios

1026
 1027
 1028
 1029
 1030
 1031
 1032
 1033
 1034
 1035
 1036
 1037
 1038
 1039
 1040
 1041
 1042
 1043
 1044
 1045
 1046
 1047

Settings	Document Languages		Query Languages					Avg.
	Hop-1	Hop-2	En	Fr	Ru	Ar	Zh	
Monolingual	Q	Q	51.68	41.15	48.02	45.67	37.38	44.81
	\overline{Q}	Q	52.72	42.60	47.22	44.32	37.39	44.85
Multilingual	Q	\overline{Q}	39.69	34.27	37.73	34.01	30.59	35.25
	\overline{Q}	\overline{Q}	40.11	34.56	36.49	30.08	27.23	33.69

1055
 1056 Table 5: Multilingual two-hop QA performance for Qwen-3-8B. We report the F1 score token ac-
 1057 curacy. Q/\overline{Q} denotes whether the languages of documents are the same or different from the query
 1058 language. For languages of documents that differ from the query, we report their average perfor-
 1059 mance.

1060
 1061
 1062
 1063
 1064
 1065
 1066
 1067
 1068
 1069
 1070
 1071
 1072
 1073
 1074
 1075
 1076
 1077
 1078
 1079