

Do Language Models Reason Across Languages?

Yan Meng Wafaa Mohammed Christof Monz

Language Technology Lab
University of Amsterdam
Science Park 904, The Netherlands
{y.meng}@uva.nl

Abstract

Real-world data sources are inherently multilingual, which raises the question *whether* language models can synthesize information across languages. In this paper, we introduce a simple two-hop question answering setting, where answering a question requires making inferences over two documents in different languages. We find that language models are more sensitive to language variation in answer-span documents than in those providing bridging information, despite the equal importance of both documents for answering a question. Under a step-by-step sub-question evaluation, we further show that in up to 33% of multilingual cases, models fail to infer the bridging information in the first step yet still answer the overall question correctly. This indicates that reasoning in language models, especially in multilingual settings, does not follow a faithful step-by-step decomposition. Moreover, language models also struggle to compose intermediate results: in up to 18% of cases, they answer both sub-questions correctly but fail to derive the correct final answer. To examine whether explicit decomposition can help, we evaluate a three-stage SUBQ prompting setup that provides the relevant intermediate sub-questions during inference. This improves two-hop QA performance on both the compositional failure cases and the full evaluation set.

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 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 distributed across languages. Enhancing multilingual reasoning is therefore essential for building

globally reliable AI systems. This requires models not only to comprehend information in multiple languages but also to integrate knowledge across them (Ghosh et al., 2025).

Recent works (Qin et al., 2023; Huang et al., 2023; Chai et al., 2025) explore multilingual reasoning in non-linguistic domains such as mathematics and coding, which inherently require multi-step reasoning. However, their evaluation is solely based on final performance, e.g., accuracy, which cannot truly capture whether language models perform inference across multiple multilingual sources or instead rely on shortcut heuristics. In this work, we focus on a *decomposed* two-hop question answering task that requires chaining information in two contexts to answer a question. The controlled decomposition of this task enables fine-grained evaluation of multilingual reasoning at each step.

To this end, we extend two English-language multi-hop datasets, the decomposed version of HOTPOTQA (Tang et al., 2021) and MUSIQUE (Trivedi et al., 2022), to four diverse and high-resource languages: French, Russian, Arabic, and Chinese.¹ This extension broadens the reasoning evaluation across languages but also bridges the gap between English-centric two-hop reasoning and multilingual NLP. 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 models perform reasoning across contexts in vary-

¹Dataset available at <https://github.com/vivian-my/X-Hop>.

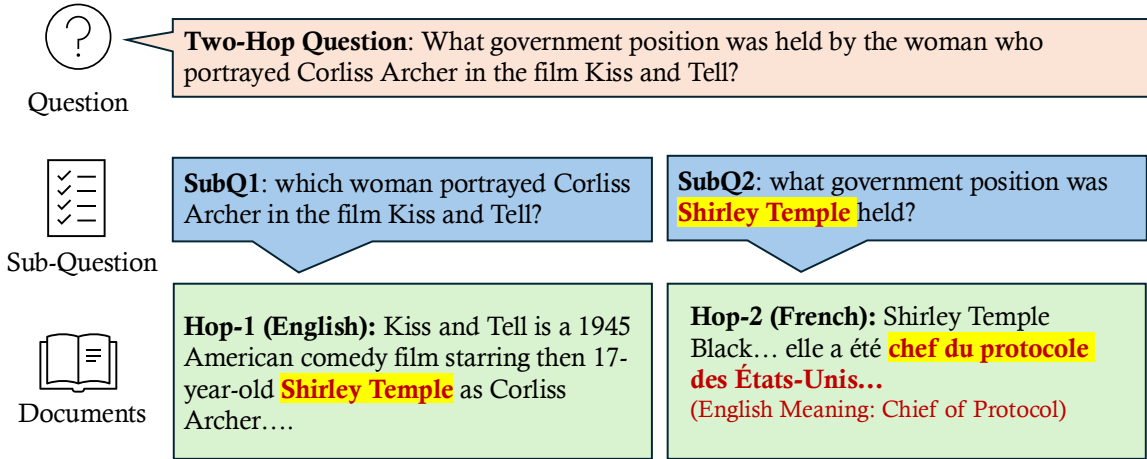


Figure 1: Example of *multilingual two-hop QA*. In Section 2, we evaluate multilingual two-step reasoning performance with the two-hop question and the corresponding Hop-1 and Hop-2 document. In Section 3, we conduct a sub-question evaluation to disentangle the two-step reasoning mechanism: SubQ1 infers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

ing languages and whether they reason in a step-by-step manner, commonly analogous to how humans solve problems via sub-questions.

From a controlled evaluation of multilingual large language models on this benchmark, we observe a degradation in two-hop QA performance when varying languages for each document (see Section 2). In particular, changing the language of the answer-span (second-hop) documents leads to larger performance drops than changing the language of the bridging (first-hop) documents. This discrepancy raises the question of whether the first step of multilingual reasoning is intrinsically easier than the second, or whether models rely on other mechanisms that bypass faithful reasoning.

Subsequently, we conduct a step-level evaluation on decomposed sub-questions (see Section 3). We observe a notable unfaithfulness ratio (up to 33%) in multilingual scenarios where language models correctly answer a two-hop question but fail to answer the first step’s sub-question. This indicates that successfully performing two-hop QA does not strictly require models to explicitly solve intermediate reasoning steps.

To understand how models reach the correct answer without identifying explicit 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. Moreover, the performance of unfaithful two-hop cases drops sharply when distractors with topically related bridging information are introduced. This

shows that language models perform multilingual two-hop reasoning, but this reasoning does not indicate explicit 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. To resolve this issue, we examine the impact of presenting explicit reasoning information. We first show that standard chain-of-thought prompting (Wei et al., 2022), which requests models to “think step-by-step”, only partially alleviates compositional failure. Building on the decomposed sub-questions, we introduce a three-stage SUBQ prompting method, where each sub-question is explicitly provided to guide the final answer generation. This approach substantially reduces compositional failures, improving averaged token-level F1 from 7.16% to 39.77%. These results highlight question decomposition as a promising direction for enhancing multilingual multi-step reasoning.

2 Multilingual Two-Hop QA

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

Settings	Document Languages		Query Languages (L_q)					Avg.
	Hop-1	Hop-2	En	Fr	Ru	Ar	Zh	
Monolingual	$= L_q$	$= L_q$	56.08	49.04	43.48	43.70	32.25	44.91
	$\neq L_q$	$= L_q$	55.49	49.19	43.12	42.92	35.41	45.22
Multilingual	$= L_q$	$\neq L_q$	48.58	45.67	40.01	40.19	36.13	42.12
	$\neq L_q$	$\neq L_q$	48.16	44.64	41.37	40.50	40.33	43.00

Table 1: Multilingual two-hop QA performance for Gemma-3-27B-Instruct. We report token-level F1 scores based on exact token matching. $= L_q$ or $\neq L_q$ denote whether the languages of documents are the same or different from the query language. For non-matching documents, we report the average across the four non-query languages. Full results are in Appendix G. Results for additional models, including Llama-3.1, Qwen-3, and Aya-Expanse, are reported in Appendix F.

Hop-1 as first-hop documents containing bridge entities and Hop-2 as second-hop documents with target answers. In particular, a bridge entity must first be identified to infer the final answer in the second-hop documents. This task evaluates models’ multilingual understanding and reasoning ability to integrate information across different languages.

2.1 Experimental Setup

Datasets. We evaluate on the extended decomposed HOTPOTQA dataset (Tang et al., 2021), derived from Yang et al. (2018), which comprises 1,000 English two-hop questions with corresponding decomposed sub-questions; we use SubQ1 and SubQ2 to denote the first and second single-hop sub-questions. To probe multi-hop reasoning at varying depths, we additionally use MUSIQUE (Trivedi et al., 2022), whose dev set contains 2,417 compositional questions spanning 2-, 3-, and 4-hop reasoning chains.

Dataset Filter. To ensure models rely on true compositional reasoning rather than memorized pre-training knowledge, we filter both datasets as follows. For a k -hop question with k gold documents $\{d_{\text{hop}_1}, \dots, d_{\text{hop}_k}\}$, we discard any data instance that the experimental models can answer correctly from any subset of these documents, i.e., any combination excluding at least one required document. This ensures that deriving the correct answer strictly requires the complete k -document chain. After filtering, the evaluation corpus contains 1,312 examples: 803 two-hop, 327 three-hop, and 182 four-hop.

Dataset Translation. We automatically translate the filtered English datasets into four high-resource languages with varying language families and writ-

Setting	Number of Hops		
	2	3	4
Monolingual F1	56.1	48.8	40.5
Cross-lingual F1	50.8	43.2	30.9
Δ F1	-5.3	-5.6	-9.6

Table 2: Token-level F1 scores of Gemma-3-27B-Instruct as the number of reasoning hops increases. The query here is English. In the monolingual setting, all supporting documents are English. In the cross-lingual setting, at least one supporting document is not English (averaged over Chinese, French, Arabic, and Russian).

ten 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.² To ensure translation quality, we conducted human evaluations on a subset of translation examples (see Appendix B).

Models. We experiment with the Gemma-3-Instruct (Kamath et al., 2025) model of size 27B due to its strong multilingual ability, which supports over 140 languages. To show the generalisability of our findings, we report results from additional multilingual models, including Qwen-3-8B, Llama-3.1-8B-Instruct, and Aya-Expanse-8B, in Appendix F. Following Modarressi et al. (2025), we adopt greedy decoding for question answering tasks. 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 D.

²Sampling parameters for translation: temperature = 1.0 and Top- p = 1.0.

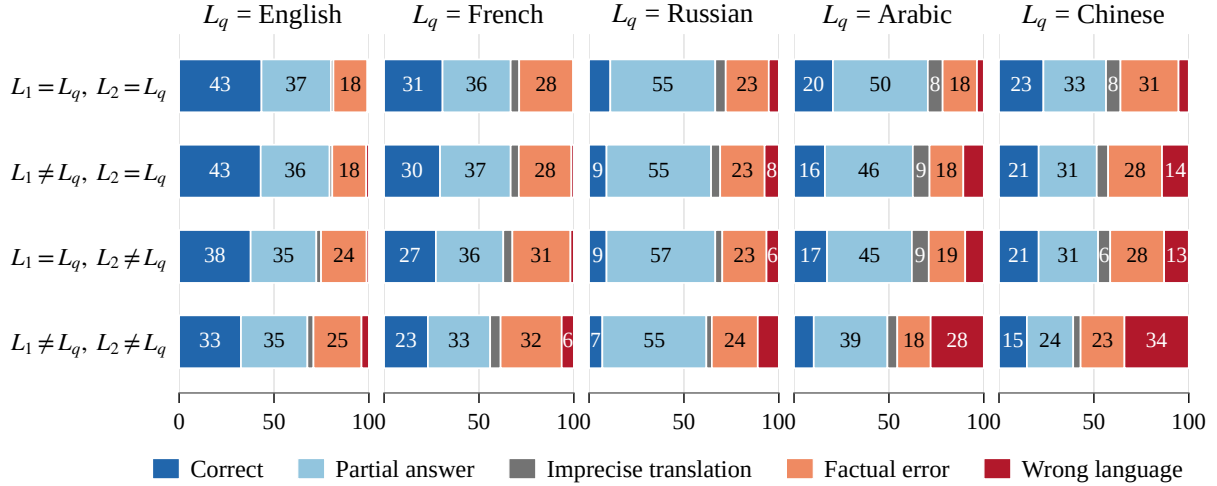


Figure 2: Error analysis of two-hop QA predictions from Gemma-3-27B-Instruct. L_q , L_1 , and L_2 denote the languages of the query, Hop-1 document, and Hop-2 document, respectively.

2.2 Results

Cross-lingual degradation persists across reasoning depths. Table 1 reports multilingual two-hop QA performance under different query languages. Across the evaluated multilingual language models, performance mostly decreases in cross-lingual settings where one or both document languages differ from the query language. This degradation is not limited to two-hop reasoning: Table 2 shows that cross-lingual settings consistently underperform monolingual settings as the number of reasoning hops increases.

Reasoning-related errors dominate cross-lingual degradation. We define a fine-grained taxonomy of errors in two-hop QA predictions (Appendix E), including partial answers, imprecise translations, factual errors, and wrong-language responses. We annotate 500 predictions from Gemma-3-27B for each query language (2,500 predictions in total) using Claude-OPUS-4.6. Figure 2 shows that factual failures and imprecise answers are the major errors in both monolingual and cross-lingual settings. This suggests that the cross-lingual degradation primarily reflects failures in reasoning over the multilingual information, rather than errors in translating final answers. At the same time, the failure modes vary across query languages. For English and French queries, changing to cross-lingual contexts mainly increases factual failures. For Arabic, Chinese, and Russian queries, wrong language predictions become more frequent in cross-lingual settings, although reasoning-related errors remain the primary source.

Hop-2 language matters more than Hop-1. Table 1 shows that performance is substantially more sensitive to the language of the answer-span document (Hop-2) than to that of the bridging document (Hop-1). Across all models, changing the Hop-1 language yields only a marginal average change relative to the monolingual setting (-0.83 F1), while changing the Hop-2 language leads to a much larger average drop of -8.49 F1. This pattern holds across all models, although the magnitude varies considerably, from -2.79 F1 for Gemma-3-27B to -13.43 F1 for Aya-Expansive-8B. To examine whether this degradation relates to linguistic similarity, we compute Pearson and Spearman correlations between QA performance and the linguistic similarity between the question and each document.³ Table 3 shows that final performance has a strong correlation with the linguistic similarity between the question and the Hop-2 document.

3 Multilingual Reasoning Decomposition

In this section, we disentangle the two-hop reasoning process by two single steps through our translated subset of decomposed sub-questions from HOTPOTQA. 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 Hop-1 documents, and SubQ2 as the second-step sub-questions that

³Details of the linguistic similarity calculation are provided in Appendix C

Correlation	En		Fr		Ru		Ar		Zh	
	Hop-1	Hop-2	Hop-1	Hop-2	Hop-1	Hop-2	Hop-1	Hop-2	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 3: Pearson/Spearman correlations between two-hop QA performance and the linguistic similarity for Gemma-3-27B-Instruct. Overall, linguistic similarity between two-hop questions and Hop-2 documents has a strong correlation with the performance.

retrieve final answers from 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).

Unfaithfulness. Models correctly answer 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 models are faithful to step-by-step sub-question reasoning.

Compositional Failure. Models incorrectly answer 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 language models’ compositional reasoning.

3.2 Unfaithfulness

Table 4a 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 first-step unfaithfulness rates of 25% and 33%, respectively. To further probe how language models are 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 documents.

3.2.1 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 and 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 ones⁴ to show shifts in context attribution.

Results. Figure 3 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 model 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 scores for unfaithful cases indicate the model relies on a more nuanced intrinsic reasoning mechanism that extends beyond what can be revealed through ex-

⁴Two-hop, SubQ1, and SubQ2 all correct

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 4: The decomposed sub-questions evaluation for Gemma-3-27B-Instruct. We report the average ratio for multilingual settings.

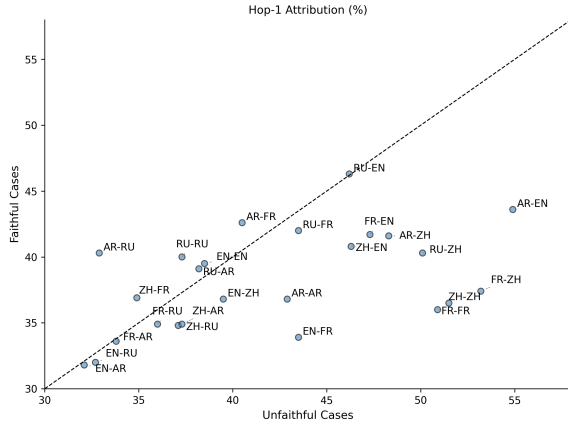


Figure 3: Context attribution scores for faithful and unfaithful cases. The two-hop query is in English, and Lang1-Lang2 (e.g., EN-ZH) indicates the languages of Hop-1 and Hop-2 documents.

licit 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 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.

3.2.2 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 shortcutting 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 4 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 models to discriminate true bridging evidence from semantically proximate noise and highlighting the central role of Hop-1 in two-step reasoning. Moreover, changing the document order further degrades performance: presenting the Hop-1 and Hop-2 documents in re-

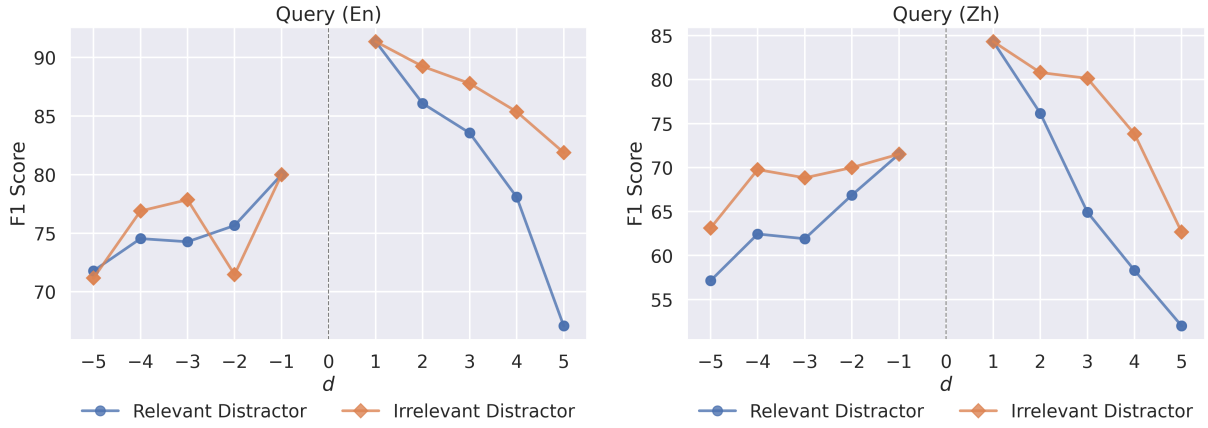


Figure 4: The impact of inserting relevant and irrelevant distractors between Hop-1 and Hop-2 documents. A distance of d corresponds to $(|d| - 1)$ distractors between the two hops. Positive d means Hop-1 precedes Hop-2, while a negative sign means the reverse. We report the average token-level F1 scores for every unfaithful multilingual case for each query language.

verse order is consistently worse than the original sequence, echoing premise-order effects in LLMs (Chen et al., 2024b; 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 4b shows compositional 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 language models with (i) self-generated chain-of-thoughts (CoT) or (ii) decomposed step-by-step sub-questions, can reduce compositional failure.

Zero-shot CoT. Following Kojima et al. (2022), we instruct the model to “*Think Step-by-Step*” and generate its own reasoning chains for answering two-hop questions. Prompt templates are in Appendix Figure 8. The language of zero-shot instruction is aligned with the two-hop question, and we do not control the reasoning language here.

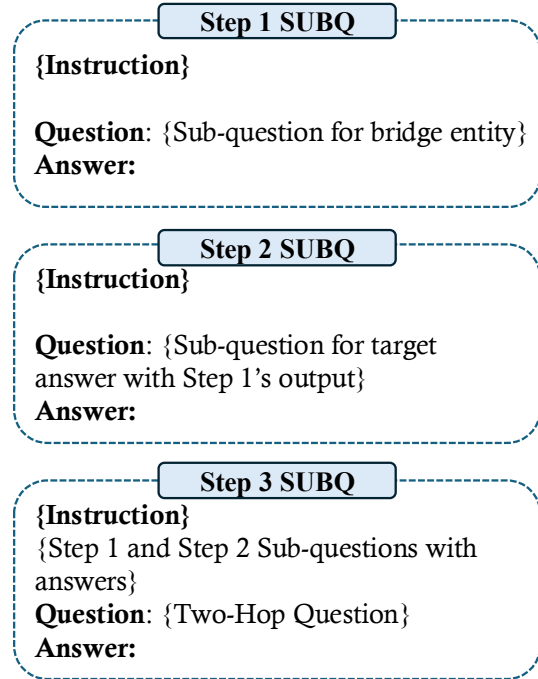


Figure 5: Three-step of SUBQ Prompting.

SUBQ Prompt. We introduce a three-stage, step-by-step prompting technique with decomposed sub-questions. As illustrated in Figure 5, 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.

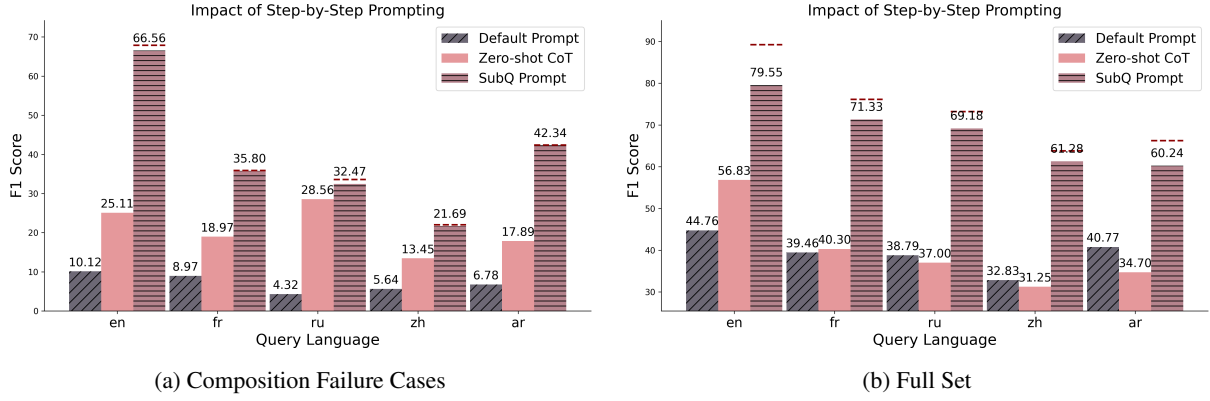


Figure 6: Token-level F1 scores for multilingual two-hop QA for different prompting strategies, evaluated on composition failure cases and the full dataset. Red dashed lines above the SubQ Prompt bars show the performance when using ground-truth answers for sub-questions.

Notably, SUBQ is not intended as a deployable method, as it builds on gold decomposed sub-questions. Instead, it serves as a diagnostic upper bound for estimating how much cross-lingual degradation can be attributed to failures in question decomposition and intermediate reasoning. Future work can investigate automatic decomposition methods to separate from the oracle supervision.

3.3.1 Results.

Figure 6 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 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.

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.

4 Related Work

In this section, we review prior work relevant to our study, focusing on multilingual capability and reasoning capability.

Multilingual Capabilities. Multilingual capabilities have been extensively studied across downstream tasks, including machine translation (Stap et al., 2023; Meng and Monz, 2024), question answering (Ponti et al., 2020; Hendrycks et al., 2021), and mathematical reasoning (Shi et al., 2023; Qin et al., 2023; Chen et al., 2024a; Ghosh et al., 2025). Among these, multilingual mathematical reasoning is closely related to our study, but primarily examines reasoning over numerical operations rather than linguistic information. Prior work has also used linguistic benchmarks, such as XCOPA (Ponti et al., 2020) and translated versions of MMLU (Hendrycks et al., 2021), to evaluate multilingual reasoning (Chua et al., 2024). However, XCOPA typically requires only a single inference step, while MMLU does not provide explicit control over the intermediate steps involved in reaching an answer. In contrast, we study multilingual multi-hop question answering in a controlled setting, disentangling the language used at each reasoning step to directly examine models’ ability to integrate linguistic information across languages.

Intrinsic Reasoning. Intrinsic reasoning investigates whether LLMs are capable of reasoning without explicit prompting (Wang and 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 occurs 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 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., 2024b; 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. Tang et al. (2021) used decomposed sub-questions to evaluate multi-hop QA, and we extend this evaluation to cross-lingual settings.

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 current strong multilingual language models in two-step reasoning when information is distributed across languages. Specifically, we find that lan-

guage models are more sensitive to cross-lingual variation in answer-span documents than in bridging documents, despite both being equally essential for deriving the final answer. Through both explicit and implicit analyses, we show that models rely on more nuanced intrinsic reasoning mechanisms that extend beyond what can be revealed by explicit step-by-step evaluation. 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.

Limitations

We acknowledge several limitations in our work. First, this paper primarily analyzes how language models reason across languages. The proposed SUBQ method is intended as a diagnostic tool for examining whether LLMs benefit from structured, step-by-step reasoning, rather than as a directly deployable method. Its applicability is limited because our setting provides oracle decomposed sub-questions, whereas real-world settings require an additional question-decomposition process. Second, our translated dataset focuses on four high-resource languages and does not include low-resource languages. This design choice aims to mitigate the impact of training data scarcity on models' cross-lingual reasoning performance. Although these four high-resource languages still have less pre-training data compared to English, we believe they provide a sufficiently controlled setting to analyze cross-lingual reasoning without the extreme data imbalance in low-resource languages.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080. We would like to thank Vlad Niculae, David Stap, and Sergey Troshin for their useful suggestions. We would also like to thank the reviewers for their feedback.

References

- Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthoran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. [Chain-of-thought reasoning in the wild is not always faithful](#). In *Workshop on Reasoning and Planning for Large Language Models*.
- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. [Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.
- Linzhang Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2025. [Xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:23550–23558.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024a. [Breaking language barriers in multilingual mathematical reasoning: Insights and observations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.
- Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. 2024b. [Premise order matters in reasoning with large language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6596–6620. PMLR.
- Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2024. [Crosslingual capabilities and knowledge barriers in multilingual large language models](#). *ArXiv*, abs/2406.16135.
- Ian Covert, Scott M. Lundberg, and Su-In Lee. 2021. [Explaining by removing: A unified framework for model explanation](#). *J. Mach. Learn. Res.*, 22:209:1–209:90.
- Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. [A survey of multilingual reasoning in language models](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I. Jordan, and Stuart Russell. 2025. [How do LLMs perform two-hop reasoning in context?](#) *ArXiv*, abs/2502.13913.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. [Exploring human-like translation strategy with large language models](#). *Transactions of the Association for Computational Linguistics*, 12:229–246.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2025. [Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5165–5180, Albuquerque, New Mexico. Association for Computational Linguistics.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. [Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.
- Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. [Gemma 3 technical report](#). *ArXiv*, abs/2503.19786.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). *ArXiv*, abs/2205.11916.
- Kenneth J. Kurtz, Dedre Gentner, and Virginia Gunn. 1999. [Chapter 4 - reasoning1](#). In Benjamin Martin Bly and David E. Rumelhart, editors, *Cognitive Science*, pages 145–200. Academic Press, San Diego.
- Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. [Understanding neural networks through representation erasure](#). *CoRR*, abs/1612.08220.
- Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025. [Logical reasoning in large language models: A survey](#). *Preprint*, arXiv:2502.09100.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. [Faithful chain-of-thought reasoning](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.
- Yan Meng and Christof Monz. 2024. [Disentangling the roles of target-side transfer and regularization in multilingual machine translation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1828–1840, St. Julian’s, Malta. Association for Computational Linguistics.
- Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. [Nolima: Long-context evaluation beyond literal matching](#). In *Forty-second International Conference on Machine Learning*.
- Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- 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.
- Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2025. [On the consistency of multilingual context utilization in retrieval-augmented generation](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 199–225, Suzhuo, China. Association for Computational Linguistics.
- Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. [Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. [Language models are multilingual chain-of-thought reasoners](#). In *The Eleventh International Conference on Learning Representations*.
- David Stap, Vlad Niculae, and Christof Monz. 2023. [Viewing knowledge transfer in multilingual machine translation through a representational lens](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14973–14987, Singapore. 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.
- 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.
- Xuezhi Wang and Denny Zhou. 2024. [Chain-of-thought reasoning without prompting](#). *ArXiv*, abs/2402.10200.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). *ArXiv*, abs/2201.11903.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. [Do large language models latently perform multi-hop reasoning?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Sangwon Yu, Ik-hwan Kim, Jongyoon Song, Saehyung Lee, Junsung Park, and Sungroh Yoon. 2025. [Unleashing multi-hop reasoning potential in large language models through repetition of misordered context](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6435–6455, Albuquerque, New Mexico. Association for Computational Linguistics.

A Use of Large Language Models

We use large language models (LLMs) in three aspects: First, we use GPT-4o-mini to translate the dataset into the target languages. Second, we use ChatGPT for grammar-level polishing of the manuscript. Third, we use Claude-OPUS-4.6 for error annotation. We do not use LLMs for research ideation, experimental design, or generating substantive scientific content.

B Translation Quality Evaluation

We evaluate translation quality using both automatic and human evaluation. For automatic evaluation, we report reference-free COMETKiwi⁵ scores (Rei et al., 2022). For human evaluation, we randomly sample 20% of the translated dataset for each language. We invite PhD students at the University of Amsterdam to rate the translated dataset. They are native speakers of French, Chinese, Arabic, and Russian but also speak English fluently. They volunteered to participate in the translation rating task and therefore received no compensation. They were informed of the purpose of the translation-rating task and were instructed to evaluate the dataset using the four-point scale described below. Results are in Table 5.

- **3:** The translation preserves the full meaning of the English source and contains no grammatical errors.
- **2:** The translation preserves most of the meaning of the English source, with only minor grammatical errors.
- **1:** The translation preserves only part of the source meaning and may contain fluency or grammatical issues.
- **0:** The translation preserves little or none of the source meaning and is hard to understand.

C Linguistic Similarity Calculation

We compute linguistic similarity using subword vocabulary overlap on the NTREX multi-parallel corpus,⁶ following Qi et al. (2025). NTREX covers 128 languages and provides parallel text for estimating vocabulary overlap across languages.

⁵<https://huggingface.co/Unbabel/wmt22-cometkiwi-da>

⁶<https://github.com/MicrosoftTranslator/NTREX>

For two languages l_1 and l_2 , we define the overlap score as:

$$\text{Overlap}(l_1, l_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|}, \quad (1)$$

where V_1 and V_2 denote the subword vocabularies induced from languages l_1 and l_2 , respectively.

D Prompts

Figure 7 and 8 present prompt templates used for different experiments in this paper. In all settings, the instruction language is matched to the query language, with instructions translated by native speakers for each target language.

E Error Analysis

- **Correct:** The prediction exactly matches the gold answer (i.e., Token-level F1 is 1).
- **Partial Answer:** The prediction identifies the correct answer but is incomplete or overly specific. For example, given the gold answer “Erick Sermon, Redman and Keith Murray,” predicting only “Erick Sermon”; similarly, predicting “1950s” for “From the 1950s to the 1970s”.
- **Imprecise Translation:** The prediction identifies the correct underlying answer but contains translation errors in the target language.
- **Factual Errors:** The prediction is in the correct language but gives factually incorrect information.
- **Wrong Language:** The prediction is in a language different from the query language, regardless of whether the underlying answer is factually correct.

F Additional Model Results

Tables 6, 7, and 8 report additional results for Qwen-3-8B, Aya-Expanse-8B and Llama-3.1-8B-Instruct.

G Full Multilingual Two-hop QA Results

Figure 9 presents the full multilingual two-hop QA results for Gemma-3-27B-Instruct across five query languages under different multilingual controlled settings.

Metric	French (Fr)	Russian (Ru)	Chinese (Zh)	Arabic (Ar)
COMETKiwi	86.14	83.42	82.89	80.43
Human	2.60	2.40	2.50	2.60

Table 5: Translation quality of the multilingual two-hop QA dataset. We report reference-free COMETKiwi scores and average human ratings on a 0–3 scale. Both automatic and human evaluations indicate that the translations largely preserve the meaning of the English source questions and documents.

Monolingual Two-Hop Prompting

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: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Article:

Hop-1 Doc: Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer....

Hop-2 Doc: Shirley Temple Black was an American actress....She was named United States ambassador to Ghana and to Czechoslovakia and also served as Chief of Protocol of the United States.

Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Answer:

(a) Monolingual setting: English query, English Hop-1 document, and English Hop-2 document.

Multilingual Two-Hop Prompting

您将收到一篇文章和一个问题。请根据给定的文章尽可能简洁地回答问题，尽可能使用一个短语或句子。请勿提供任何解释。

问题：在电影《吻与告白》中饰演科尔利斯·阿彻的女性担任了什么政府职位？

文章：

Hop-1 Doc: Kiss and Tell est une comédie américaine de 1945 mettant en vedette Shirley Temple...

Hop-2 Doc: وديبلوماسية، وكانت شيرلي تيمبل بلاك (23 أبريل 1928 - 10 فبراير 2014) كانت ممثلة أمريكية، مغنية، راقصة، سيدة أعمال، الرقم واحد في شباك التذاكر في هوليوود كممثلة أطفال من 1935 إلى 1938. كبالغة، تم تعيينها سفيرة الولايات المتحدة لدى غانا وتشيكوسلوفاكيا وكذلك خدمت كرئيسة للبروتوكول في الولايات المتحدة.

问题：在电影《吻与告白》中饰演科尔利斯·阿彻的女性担任了什么政府职位？

答案：

(b) Cross-lingual setting: Chinese query, French Hop-1 document, and Arabic Hop-2 document.

Figure 7: Default multilingual two-hop QA prompt templates. The labels **Hop-1 Doc** and **Hop-2 Doc** are shown only for illustration and are not included in the actual prompt.

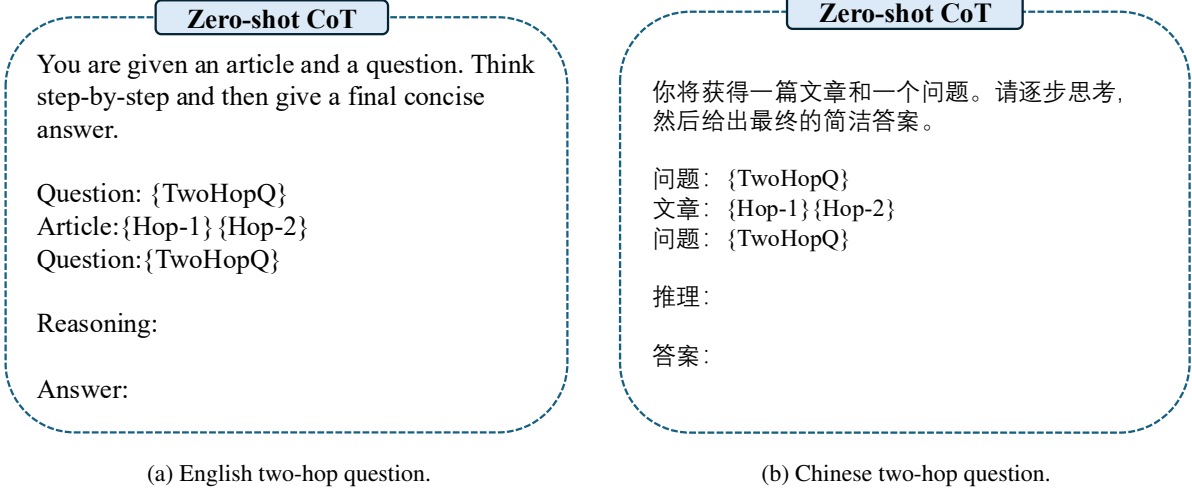


Figure 8: Zero-shot CoT prompt templates. We append “think step-by-step” to encourage the model to generate an intermediate reasoning chain before answering.

Setting	Document Languages		Query Languages (L_q)					Avg.
	Hop-1	Hop-2	En	Fr	Ru	Ar	Zh	
Monolingual	$= L_q$	$= L_q$	52.59	45.64	38.62	37.73	27.47	40.41
	$\neq L_q$	$= L_q$	52.99	44.56	38.63	37.23	25.84	39.85
Cross-lingual	$= L_q$	$\neq L_q$	41.80	36.30	29.19	27.25	23.89	31.69
	$\neq L_q$	$\neq L_q$	39.78	34.79	28.40	26.89	24.12	30.80

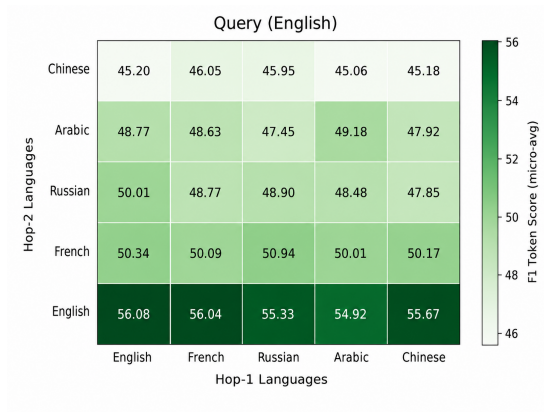
Table 6: Multilingual two-hop QA performance for Qwen-3-8B. We report token-level F1. $= L_q$ and $\neq L_q$ indicate whether the document language matches or differs from the query language. For non-matching documents, we report the average across the four non-query languages.

Setting	Document Languages		Query Languages (L_q)					Avg.
	Hop-1	Hop-2	En	Fr	Ru	Ar	Zh	
Monolingual	$= L_q$	$= L_q$	60.06	43.30	34.91	42.08	36.07	43.29
	$\neq L_q$	$= L_q$	58.08	43.41	34.62	41.56	34.71	42.47
Cross-lingual	$= L_q$	$\neq L_q$	25.27	33.12	31.86	36.35	22.71	29.86
	$\neq L_q$	$\neq L_q$	19.94	31.88	31.41	34.89	18.82	27.39

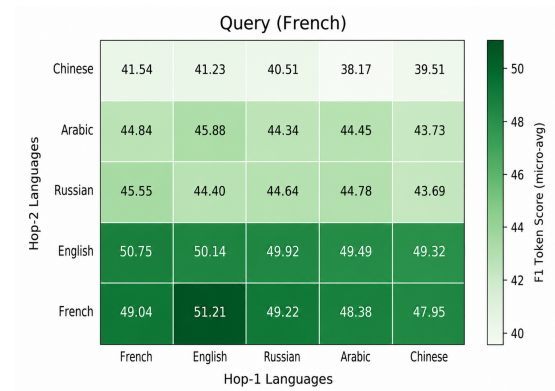
Table 7: Multilingual two-hop QA performance for Aya-Expans-8B. We report token-level F1. $= L_q$ and $\neq L_q$ indicate whether the document language matches or differs from the query language. For non-matching documents, we report the average across the four non-query languages.

Setting	Document Languages		Query Languages (L_q)					Avg.
	Hop-1	Hop-2	En	Fr	Ru	Ar	Zh	
Monolingual	$= L_q$	$= L_q$	46.64	39.07	29.24	28.30	23.24	33.30
	$\neq L_q$	$= L_q$	43.91	36.21	28.60	24.62	21.90	31.05
Cross-lingual	$= L_q$	$\neq L_q$	35.49	30.55	22.10	18.64	14.67	24.29
	$\neq L_q$	$\neq L_q$	30.10	28.57	22.27	18.86	16.47	23.25

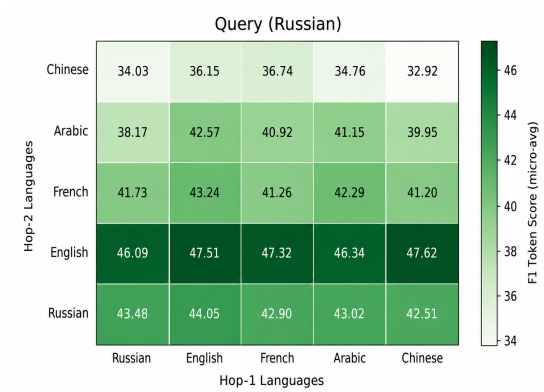
Table 8: Multilingual two-hop QA performance for Llama-3.1-8B-Instruct. We report token-level F1. $= L_q$ and $\neq L_q$ indicate whether the document language matches or differs from the query language. For non-matching documents, we report the average across the four non-query languages.



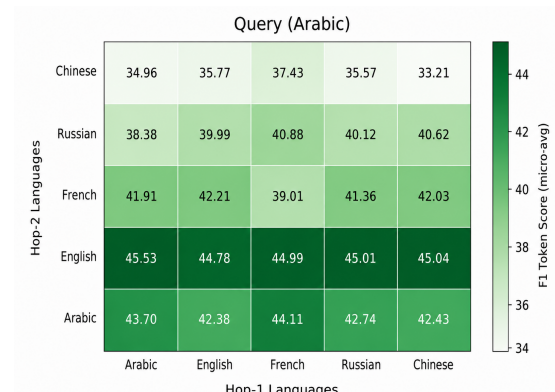
(a) English



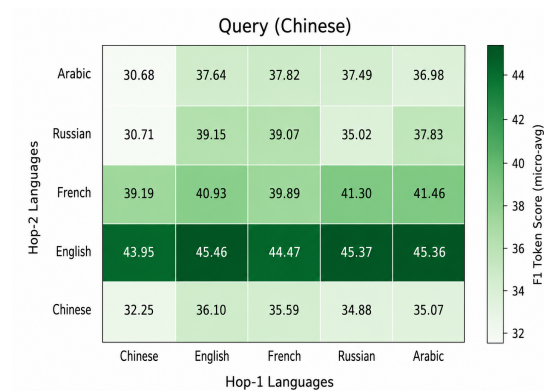
(b) French



(c) Russian



(d) Arabic



(e) Chinese

Figure 9: Full multilingual two-hop QA performance for Gemma-3-27B-Instruct across query languages. Each panel fixes the query language and varies the languages of the Hop-1 and Hop-2 evidence documents.