MULTITAT: Benchmarking Multilingual Table-and-Text Question Answering

Anonymous ACL submission

Abstract

Question answering on the hybrid context of tables and text (TATQA) is a critical task, with broad applications in data-intensive domains. However, existing TATQA datasets are limited to English, leading to several drawbacks: (i) They overlook the challenges of multilingual TAT-QA and cannot assess performance in the multilingual setting. (ii) They do not reflect real-world multilingual scenarios where tables and texts frequently appear in non-English languages. To address the limitations, we propose the first multilingual TATQA dataset (MULTITAT). Specifically, we sample data from 3 mainstream TATQA datasets and translate it into 10 diverse languages. To align the model TATQA capabilities in English with other languages, we develop a baseline, OURS. Experimental results reveal that the performance on non-English data in MULTITAT drops by an average of 19.4% compared to English, proving the necessity of MULTITAT. We further analyze the reasons for this performance gap. Furthermore, OURS outperforms other baselines by an average of 3.3, demonstrating its effectiveness[1].

1 Introduction

Question answering over the hybrid context of tabular and textual data (TATQA) is an important task (Chen et al., 2020), which is widely used in data-intensive fields, such as finance and science, gaining increasing attention (Chen et al., 2021; Auer et al., 2023). Enhancing the TATQA capabilities of models can significantly aid in extracting useful information from hybrid data. The heterogeneous evidence brings challenges to the TATQA task since it requires the model to link the relevant information in the table or text according to the entities in the question (Feng et al., 2022; Lei et al., 2022; Wang et al., 2022).

[1]Our data will be released upon acceptance.

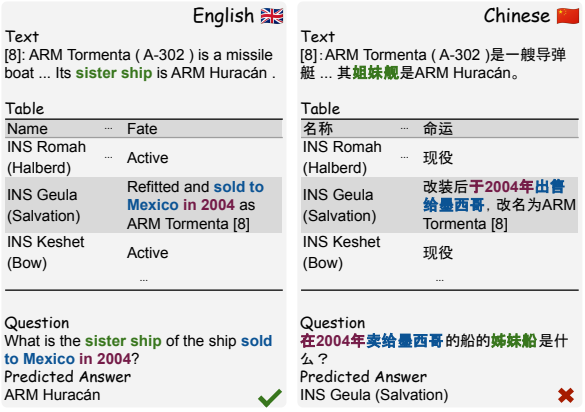


Figure 1: Comparison of the English and Chinese examples in MULTITAT. Entities with the same color annotation represent corresponding entity information. In Chinese, the richness of lexical expressions makes it more challenging for the model to link relevant information, leading to the incorrect predicted answer.

To evaluate the model capabilities on the TATQA task, several datasets are proposed (Li et al., 2021; Chen et al., 2021; Zhao et al., 2024b). For example, HybridQA (Chen et al., 2020), TAT-QA (Zhu et al., 2021), and SciTAT (Zhang et al., 2024a) respectively construct English TATQA datasets in the domains of Wikipedia, finance, and science. However, these datasets focus solely on English, having the following shortcomings: (*i*) They cannot adequately assess the TATQA performance in the multilingual setting, ***overlooking the challenges of multilingual TATQA***. As shown in Figure 1, the complex lexical expressions of different languages pose challenges for models to link information across hybrid contexts (Dou et al., 2023). (*ii*) They ***create a gap with real-world multilingual scenarios***, as domains such as finance and science contain substantial amounts of non-English tables and text (Hamotskyi et al., 2024; Angulo et al., 2021; Bhagavatula et al., 2012). To address the limitations, we propose the first multilingual

TATQA benchmark, comprising parallel data in 11 diverse languages.

First, we introduce the multilingual TATQA dataset (MULTITAT). To ensure the high quality of MULTITAT, we sample data from three mainstream English TATQA datasets and employ a combination of machine translation and manual revision to translate them into 10 languages. In total, MULTITAT consists of 250 questions from 233 hybrid contexts, covering three domains: Wikipedia, finance, and science.

To enhance the performance of MULTITAT **on non-English languages**, we propose a baseline to bridge the performance gap between English and non-English on TATQA (OURS). To align the model TATQA capabilities in English with other languages, especially low-resource languages, OURS is divided into two modules: linking non-English information and reasoning in English. Specifically, OURS first identifies relevant information from tables and text according to the entities in the question through linking and then uses this information to perform reasoning in English by generating programs.

We evaluate the performance of OURS, compared with a series of baselines on MULTITAT. Experimental results indicate that the performance of non-English languages drops by an average of 19.4% compared to English on all baselines, highlighting the necessity of MULTITAT. OURS outperforms other baselines by an average of 3.3, demonstrating the effectiveness. Analysis experiments reveal that the TATQA capabilities across languages are not only influenced by resource availability but also by their specific linguistic characteristics. Error analysis shows that the performance decline in non-English TATQA is primarily due to the reduced ability to link relevant information, apply formulas, and follow instructions.

Our contributions are as follows:

1. To the best of our knowledge, we introduce the first multilingual TATQA dataset MULTITAT, which includes 11 diverse languages.

2. We propose OURS, a baseline to align the model TATQA capabilities in English to non-English languages.

3. We conduct a series of experiments, supported by empirical results and error analysis, to demonstrate the challenges of MULTITAT and provide insights for future improvements.

2 MULTITAT

The input of MULTITAT consists of a question, the hybrid context including the table and text, and the output is the answer. Additionally, we annotate the rationale, which is the reasoning process of answering the question. We refer to each question, along with its table, text, rationale, and answer, as an instance. For each instance, we annotate 11 diverse languages. We first describe the construction process of MULTITAT, which combines automatic translation with manual error correction, following previous works (Peng et al., 2024; Singh et al., 2024; Dou et al., 2023), as shown in Figure 2.

2.1 Data Preparation

We first collect English data from existing datasets and select languages to translate them.

2.1.1 Source Data Collection

We select HybridQA (Chen et al., 2020), TATQA (Zhu et al., 2021), and SciTAT (Zhang et al., 2024a) datasets from the Wikipedia, finance, and science domains as our data sources, as these three domains are the primary areas where TATQA tasks are distributed (see Appendix A). To ensure an even distribution of different answer types and answer sources in MULTITAT, we sample a total of 250 instances from the three datasets according to the proportions shown in Table 1. Among them, only 50 instances are sampled from HybridQA due to its relatively limited answer sources and types.

2.1.2 Target Language Selection

For MULTITAT, we select 11 languages, covering 8 language families: Bengali (bn), Chinese (zh), English (en), French (fr), German (de), Japanese (ja), Russian (ru), Spanish (es), Swahili (sw), Telugu (te), and Thai (th), following the previous benchmark (Shi et al., 2023). Additionally, we preserve the Arabic numerals from the original datasets across all languages to facilitate evaluation (Shi et al., 2023).

2.2 Rationale Annotation

We first demonstrate how to annotate English rationales by employing the large language model (LLM) in combination with manual refinement. We use `gpt-4o` (OpenAI et al., 2024) to complete **rationale generation** due to its strong reasoning and instruction-following capabilities. Specifically, we input the question, relevant tables and texts, and the answer into the LLM, prompting the LLM to

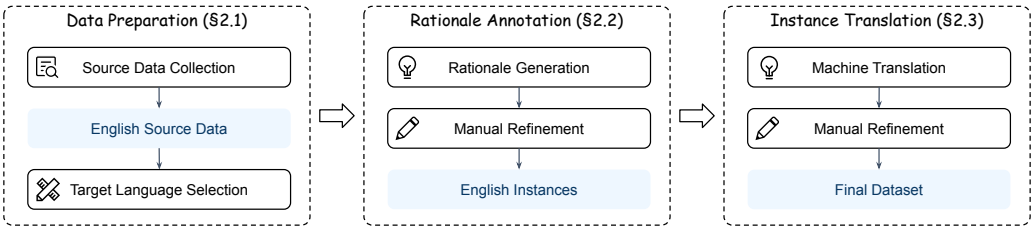


Figure 2: The process of constructing MULTITAT. The blue boxes represent the data, and the white solid boxes represent the construction steps.

Dataset	Domain	Scale	Answer Type	Answer Source			Total
				Text	Table	Hybrid	
HybridQA (Chen et al., 2020)	Wikipedia	50	Span	0	0	50	50
TAT-QA (Zhu et al., 2021)	Finance	100	Span	10	10	20	40
			Arithmetic	10	10	30	50
			Count	2	3	5	10
SciTAT (Zhang et al., 2024a)	Science	100	Span	10	20	20	50
			Arithmetic	10	20	20	50
Total	-	250	-	42	63	145	250

Table 1: The distribution of English data, including answer types and answer sources in MULTITAT, sourced from three mainstream datasets. The listed answer types are the all answer types corresponding to each dataset.

generate the corresponding rationale. Since LLMs cannot guarantee the accuracy of reasoning, we employ **manual refinement**. The annotators are instructed to evaluate the accuracy of the generated rationale and make corrections where necessary.

2.3 Instance Translation

For **machine translation**, we select `gpt-4o` because of its strong translation capabilities (Yan et al., 2024; Hu et al., 2024). Specifically, we input each instance into the LLM, with prompts to translate it into the target languages, respectively. To assess the accuracy of the translations, we use `gpt-4o` to translate the target language instances back into English, and calculate the F1 score between the back-translated version and the original English instance following previous works (Peng et al., 2024). For instances with an F1 score below 0.6, we prompt annotators to complete **manual refinement** by using Google Translation.

2.4 Quality Control

To ensure the quality of MULTITAT, we implement rigorous quality control strategies.

Competent Annotators The annotators we hire hold graduate-level degrees, are proficient in English, and are compensated with $1 per data instance. We first train the annotators to familiarize them with the annotation requirements and the use of the annotation tool (see Appendix B.1). Then, they try to annotate 20 instances, and we review their annotations, providing feedback and suggestions for revisions.

Model Evaluation To further evaluate the translation quality of MULTITAT, we employ a popular translation evaluation method (Kocmi and Federmann, 2023). The detailed evaluation method is described in Appendix C, and Table 5 presents the average translation accuracy scores. The results indicate that the average translation accuracy scores across languages in MULTITAT are above 95 out of 100, demonstrating the high translation quality.

2.5 Data Analysis

We show the data distribution of MULTITAT in Table 1. The 250 questions in MULTITAT involve 233 hybrid contexts, each of which includes 1 table and an average of 5.3 paragraphs. Each table has an average of 10.2 rows and 4.7 columns.

3 OURS

3.1 Overview

OURS is designed to address **the TATQA task under the multilingual setting**. To align the strong TATQA capabilities of models in English with non-English languages, particularly low-resource lan-

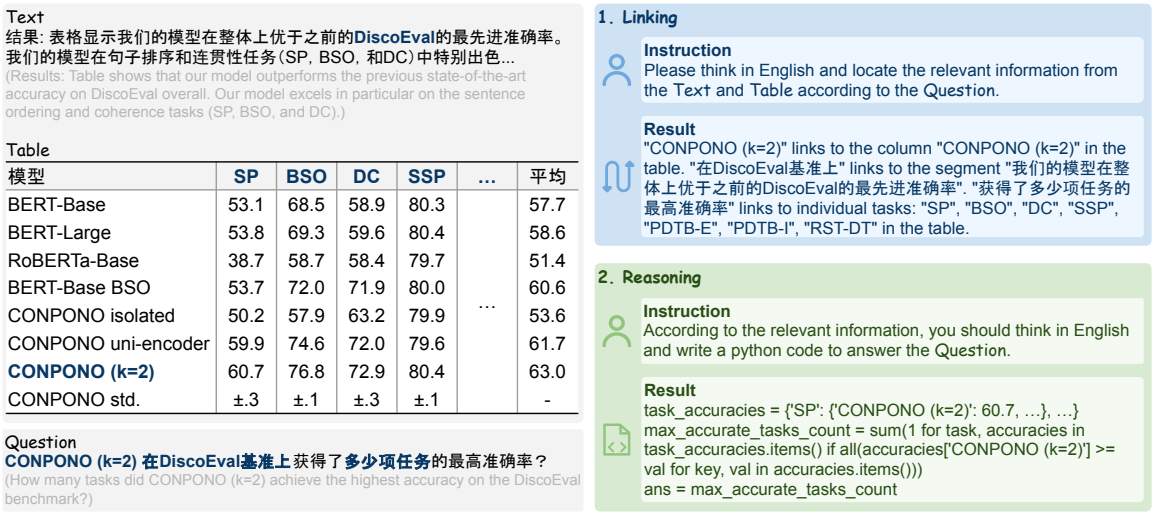


Figure 3: The overview of OURS, which includes two modules: (*i*) **Linking**: Mapping the entities in the question to the relevant information in tables or text, which are marked with **blue** in the left part. (*ii*) **Reasoning**: Generating programs to solve the question using the information. We take the Chinese TATQA input as an example, with the corresponding English text provided in (gray).

guages, OURS employs cross-lingual reasoning. To enable the model to perform English reasoning with non-English questions, tables, and text, OURS is divided into two modules: Linking and Reasoning. As shown in Figure 3, Linking is responsible for locating relevant information from tables and text in the native language based on the question, and Reasoning performs reasoning in English based on the linked information. The prompts used in OURS are provided in Appendix D.

3.2 Linking

Linking is used to map the entities in the question to the relevant information in the input text and tables so that Reasoning can directly utilize this information when generating the code. Specifically, we prompt the LLM to think in English and gradually map the relevant entities in the question to the information in the tables or text.

3.3 Reasoning

Reasoning is responsible for generating Python programs to solve the question and obtain the final answer based on the results of Linking. Considering that there are not only numbers in the answers, we also remind the LLM to note that the answers should be represented in the native language except for Arabic numerals. Since the relevant information is extracted during Linking, Reasoning can directly use English variable names to define the numerical or tabular data when generating the program.

4 Experiments

4.1 Settings

Metrics We use Exact Match (EM) and F1 score to evaluate the answers, following prior works (Chen et al., 2020; Zhu et al., 2021). EM refers to the proportion of predictions that exactly match the gold answer, and F1 measures the degree of overlap between the predicted and the gold answer in terms of their bag-of-words representation.

Models We evaluate MULTITAT using the open-source model Llama3.1-Instruct (Llama3.1) (Dubey et al., 2024) and the closed-source model `gpt-4o` (OpenAI et al., 2024). Llama3.1 is currently one of the best-performing open-source models, and `gpt-4o` is considered one of the leading closed-source models. We also evaluate MULTITAT using Qwen2.5-Instruct (Yang et al., 2024) in Appendix E.1.

Baselines We compare OURS with the following baselines with three-shot prompts, following previous works (Shi et al., 2023; Li et al., 2024).

- Native-CoT: solving the question using CoT (Wei et al., 2022) in the native language
- En-CoT: solving the question using CoT in English
- Native-PoT: prompting the LLM to generate code in the native language (Gao et al., 2023; Chen et al., 2023)
- En-PoT: prompting the LLM to generate code in English

Model	Method	bn	de	en	es	fr	ja	ru	sw	te	th	zh	Avg.
Llama3.1-8B	Native-CoT	11.2	14.0	20.8	12.8	8.0	13.2	15.2	9.2	12.4	12.0	13.6	12.9
	En-CoT	10.8	14.6	20.8	12.4	8.4	13.2	15.2	9.2	12.0	12.0	13.6	12.9
	Native-PoT	18.0	18.4	21.2	22.8	18.4	19.6	22.8	17.2	6.8	21.2	19.6	18.7
	En-PoT	13.6	12.8	21.2	20.8	14.4	20.8	20.0	10.4	7.6	19.2	19.6	16.6
	Three-Agent	10.0	16.0	21.6	20.8	15.6	13.6	12.0	13.2	9.2	15.2	18.4	15.1
	OURS	**20.0**	**22.4**	**27.6**	**25.6**	**20.0**	**25.6**	**25.2**	**17.2**	**14.4**	**22.8**	**23.6**	**22.2**
Llama3.1-70B	Native-CoT	18.8	20.8	25.6	23.6	24.8	22.4	25.2	23.6	18.8	21.6	21.6	22.4
	En-CoT	18.4	19.6	25.6	23.6	20.0	22.0	25.2	24.0	19.6	22.4	22.0	22.0
	Native-PoT	22.8	24.4	30.4	28.4	26.4	18.4	28.0	28.4	22.0	26.0	22.0	25.2
	En-PoT	23.6	26.0	30.4	27.6	26.4	25.6	28.4	26.4	22.0	25.2	26.8	26.2
	Three-Agent	16.0	25.6	29.2	23.6	22.0	25.6	20.8	22.4	20.0	19.6	23.6	22.6
	OURS	**24.0**	**28.0**	**31.2**	**29.2**	**26.8**	**26.8**	**28.8**	**30.8**	**22.8**	**26.8**	**28.0**	**27.6**
gpt-4o	Native-CoT	21.2	27.2	31.2	26.8	23.6	19.2	24.8	24.8	26.8	26.8	24.4	24.7
	En-CoT	23.6	24.8	31.2	26.0	22.0	26.4	26.4	28.0	22.0	23.2	24.8	25.3
	Native-PoT	24.4	30.4	30.0	30.4	26.4	21.2	27.2	26.4	26.8	24.8	28.0	27.6
	En-PoT	24.0	24.4	30.0	30.0	26.4	21.2	27.2	26.4	21.2	27.2	24.4	26.2
	OURS	**30.0**	**32.4**	**35.2**	**32.4**	**29.6**	**28.8**	**31.2**	**31.2**	**30.8**	**30.4**	**30.9**	**31.1**

Model	Method	bn	de	en	es	fr	ja	ru	sw	te	th	zh	Avg.
Llama3.1-8B	Native-CoT	13.2	16.1	23.7	17.2	11.2	14.5	17.3	14.0	14.9	14.6	21.5	16.2
	En-CoT	13.4	16.6	23.7	17.9	12.4	15.2	17.8	14.0	14.9	14.9	22.7	16.7
	Native-PoT	19.1	18.9	22.8	24.2	19.3	19.9	23.1	17.8	6.9	22.4	21.7	19.6
	En-PoT	14.1	13.7	22.8	21.3	15.1	21.5	20.6	11.0	7.8	20.1	21.7	17.4
	Three-Agent	15.7	20.5	26.4	25.8	20.6	15.1	16.0	17.4	13.9	18.8	26.1	19.7
	OURS	**21.3**	**24.2**	**31.9**	**27.8**	**22.4**	**26.1**	**27.0**	**20.0**	**15.2**	**24.6**	**28.0**	**24.4**
Llama3.1-70B	Native-CoT	21.6	22.8	29.3	27.0	28.1	24.4	27.3	26.6	21.3	24.0	28.3	25.5
	En-CoT	21.6	22.4	29.3	27.9	23.6	24.7	27.7	27.3	22.3	26.3	29.4	25.7
	Native-PoT	24.8	26.2	32.9	30.6	29.0	18.7	29.4	29.9	24.0	28.4	30.0	27.0
	En-PoT	25.8	27.9	32.9	30.2	28.7	27.2	30.3	28.7	25.0	27.3	30.9	28.5
	Three-Agent	22.2	30.8	34.5	31.3	28.4	28.2	25.5	27.1	24.3	24.8	33.3	28.2
	OURS	**26.3**	**31.3**	**35.3**	**34.6**	**31.1**	**29.4**	**33.5**	**34.7**	**25.9**	**30.5**	**34.9**	**31.6**
gpt-4o	Native-CoT	27.0	33.8	38.8	36.3	30.2	21.8	31.9	31.3	31.3	30.9	38.2	31.6
	En-CoT	28.0	32.1	38.8	33.1	27.2	28.8	32.4	33.6	25.0	28.8	34.6	31.1
	Native-PoT	26.7	33.3	32.5	32.5	28.7	22.5	29.9	27.7	29.4	27.2	29.5	30.1
	En-PoT	26.2	26.8	31.3	32.5	28.7	22.5	29.9	27.7	25.0	29.0	27.2	28.0
	OURS	**32.9**	**35.5**	**38.9**	**35.7**	**32.5**	**32.1**	**33.1**	**34.0**	**34.7**	**35.1**	**34.5**	**34.7**

Table 2: EM (above) and F1 (below) of different models and baselines across languages on MULTITAT. Avg. denotes the average performance of the baseline across all languages. The best results of each model under each language are annotated in **bold**.

- Three-Agent (Fatemi and Hu, 2024) is the state-of-the-art (SOTA) method on the TAT-QA dataset. It consists of three agents: the analyst agent extracts relevant data and performs computations, and two critic agents evaluate the correctness of extraction and computation, respectively, and refine the results. Due to computational resource limitations, we do not evaluate the performance of Three-Agent on MULTITAT using gpt-4o.

We present prompts for baselines and OURS in Appendix D. Additionally, we provide results for both directly answering the question and reasoning after translating the input into English in Appendix E.2.

4.2 Main Experiments

A comparison of OURS with other baselines across different languages is presented in Table 2. We observe that: (*i*) The performance on MULTITAT in non-English languages shows an average decrease of 19.4% compared to English, underscoring the necessity of MULTITAT. (*ii*) OURS demonstrates an average improvement of 3.3 on EM and F1 over other baselines, reducing the performance gap between different languages by 23.2%, which validates the effectiveness. (*iii*) Despite these improvements, the EM and F1 of all baselines remain below 40, highlighting the challenges of MULTITAT.

Baselines (*i*) OURS consistently outperforms Three-Agent because Three-Agent is not fully suited to HybridQA, which does not require computations (Chen et al., 2020), or SciTAT, which involves complex calculations that are challenging to the inherent capabilities of models (Zhang et al., 2024a). Additionally, the performance of multi-agent declines in non-English languages (Beyer et al., 2024; Chen et al., 2024). (*ii*) The performance difference between reasoning in the native language and English is minimal. Although LLMs demonstrate stronger reasoning capabilities in English, the TATQA, compared to other tasks, relies more heavily on the capabilities of linking information, which presents greater challenges in cross-lingual reasoning (Min et al., 2019). Therefore, OURS mitigates this challenge, leading to improved performance. (*iii*) PoT consistently outperforms CoT because numerical reasoning questions constitute a significant proportion of MULTITAT (see Table 1), making PoT more suitable for solving these questions (Chen et al., 2023; Zhao et al., 2024b).

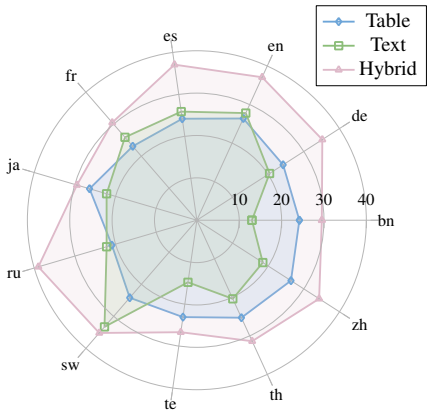


Figure 4: The EM of OURS across different answer sources on MULTITAT using Llama3.1-70B.

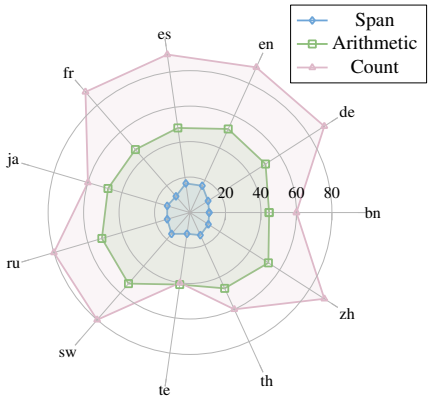


Figure 5: The EM of OURS across different answer types on MULTITAT using Llama3.1-70B.

Languages The models generally exhibit high performance on high-resource languages, such as English, German, Spanish, French, Russian, and Chinese, while their performance on low-resource languages tends to be poor. Moreover, models with stronger multilingual capabilities show smaller performance gaps across languages, with `gpt-4o` demonstrating the highest performance. This also underscores the necessity of evaluating multilingual performance on challenging tasks.

Answer Source We analyze the performance of OURS using Llama3.1-70B across different answer sources, as shown in Figure 4. The performance with other models and baselines across answer sources is provided in Appendix E.3. The results show that: (*i*) The performance of the hybrid answer source generally outperforms those with a single answer source. Since OURS, compared to other baselines (see Figure 11), enhances the links between the question and the context, integrating hybrid contextual information and alleviating the challenge. (*ii*) The performance across answer sources is influenced not only by the availability of language-specific resources but also by the characteristics of the language. For instance, languages with complex morphological structures, such as German and Russian, perform worse when the answer source is text. In contrast, Swahili shows the highest performance on text-based sources, as its simpler morphology allows for easier linking of entities in the text to those in question (Tuan Nguyen et al., 2020; Zhang et al., 2023).

Answer Type We compare the performance of OURS using Llama3.1-70B on different answer types, as shown in Figure 5. Results of other models and baselines across answer types are provided in Appendix E.4. We observe that: (*i*) The model performs best on the Count type. This is because Span answers require extracting short phrases or summarizing conclusions from tables and text, making them more sensitive to word composition and order. Additionally, Arithmetic answers involve more complex computations than Count answers. (*ii*) The model performs better on high-resource languages than low-resource languages across answer types overall. Although OURS narrows the performance gap, there remains a significant difference between high-resource and low-resource languages for all answer types.

4.3 Analysis Experiments

4.3.1 How does the Prompt Language Affect OURS?

We analyze the impact of using instructions and demonstrations in different languages on the performance of OURS, as shown in Table 3. For the multilingual demonstrations, we select one demonstration each from English, Spanish, and Chinese, as the models perform well on these three high-resource languages, which also cover two language families. The English instruction and English demonstrations are the settings of OURS used in the main experiments. The results indicate that:

(*i*) Using English instructions generally outperforms using native instructions. (*ii*) Multilingual demonstrations outperform both native language and English demonstrations, suggesting that when sufficient native demonstrations are not available on the TATQA task, using demonstrations from the same language family or high-resource languages can also enhance performance. Additionally, Swahili achieves the highest performance

Instruction	Demo	bn	de	en	es	fr	ja	ru	sw	te	th	zh	Avg.
Native	Native	20.0	28.4	28.4	29.2	29.2	27.6	27.6	**32.0**	20.4	25.2	**28.8**	27.0
	Multi	22.0	**30.0**	30.4	30.4	28.4	26.0	26.4	28.8	24.4	24.4	24.8	26.9
	En	20.8	29.2	28.4	24.8	27.2	24.0	28.4	29.2	19.6	21.2	24.4	24.9
En	Native	**27.6**	26.8	28.4	29.6	25.2	25.6	29.2	30.0	26.0	**28.0**	26.8	27.6
	Multi	26.4	27.2	30.4	**30.8**	**29.6**	**29.2**	**30.0**	30.0	**27.2**	27.2	**28.8**	**28.8**
	En	24.0	28.0	**31.2**	29.2	26.8	26.8	28.8	30.8	22.8	26.8	28.0	27.6
Instruction	**Demo**	**bn**	**de**	**en**	**es**	**fr**	**ja**	**ru**	**sw**	**te**	**th**	**zh**	**Avg.**
Native	Native	23.8	**33.9**	33.8	**35.8**	**34.0**	30.1	31.7	**35.1**	24.2	28.3	**37.4**	31.7
	Multi	24.6	32.3	**35.4**	35.0	31.6	27.6	28.8	30.7	26.3	26.7	30.7	30.0
	En	24.4	33.6	33.8	30.2	32.0	22.8	31.8	31.6	22.3	23.5	30.7	28.8
En	Native	**30.5**	30.3	33.8	32.7	29.6	28.5	33.0	33.6	28.8	**31.4**	34.1	31.5
	Multi	28.9	29.9	**35.4**	34.1	32.2	**31.7**	32.6	33.0	**29.8**	31.2	34.7	**32.1**
	En	26.3	31.3	35.3	34.6	31.1	29.4	**33.5**	34.7	25.9	30.5	34.9	31.6

Table 3: EM (above) and F1 (below) of OURS using the instructions and demonstrations of different languages on Llama3.1-70B. The best results under each language are annotated in **bold**. Demo refers to demonstrations. Multi refers to demonstrations composed of multiple languages (English, Spanish, and Chinese). Avg. denotes the average performance of the baseline across all languages.

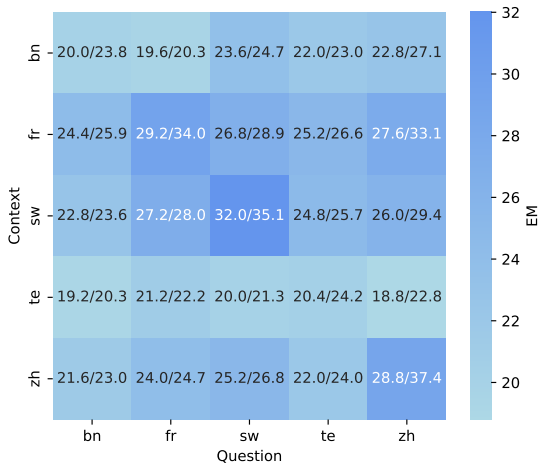


Figure 6: The EM/F1 of OURS with questions and context (table and text) of different languages on MULTI-TAT using Llama3.1-70B.

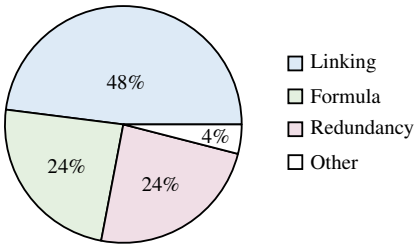


Figure 7: The error types and their proportion of non-English performance in OURS are inferior compared with English. **Linking** refers to mapping entities in the question with incorrect information in the table or text. **Formula** refers to using an incorrect formula. **Redundancy** refers to outputting irrelevant information beyond the correct answer.

when using instructions and examples in the native language, highlighting its uniqueness.

4.3.2 How does the Language Affect OURS in the Cross-lingual Setting?

We evaluate the performance of OURS in the cross-lingual setting, where the languages of the question and context are inconsistent, with results in Figure 6. We select high-resource languages (French and Chinese), and low-resource languages (Bengali, Swahili, and Telugu), covering 4 language families. Our findings include: (*i*) Generally, OURS shows improved performance when transitioning from low-resource to high-resource languages, while the opposite results in a decline. For instance, the performances on the French context with French and Chinese questions are relatively high, whereas the performances with three low-resource languages are lower. (*ii*) The model achieves the best performance when the question and context are both Swahili. This can be attributed to its relatively regular grammatical and lexical structures, which provide advantages when linking related information.

4.4 Error Analysis

We analyze the reasons for the inferior performance of OURS on non-English languages compared to English, as shown in Figure 7. Specifically, we select instances where OURS achieved an EM of 1 in English using Llama3.1-70B, but an EM of 0 in non-English languages. For each language, we randomly sample five instances, with a total of 50 errors for comparative analysis. Examples of errors corresponding to each type are provided in Appendix E.5. Below, we present a detailed discussion of each error type:

(*i*) **Linking**: Due to the relatively weaker abilities in non-English languages compared to English,

even though OURS initially prompts the model to focus on linking, the model still faces significant challenges in linking. These challenges are particularly pronounced in languages with complex orthographies, such as Japanese (with its hiragana and katakana scripts), or morphologically rich languages like French and German. (*ii*) **Formula** highlights the gap in the numerical reasoning abilities between non-English languages and English. (*iii*) **Redundancy** reflects the relatively weaker ability of instruction-following.

In summary, the inferior performance on non-English languages and the specific properties of languages leads to the lower performance of OURS on non-English languages, which also demonstrates the necessity of MULTITAT.

5 Related Works

5.1 Multilingual Datasets

To evaluate the performance of models across different languages, several multilingual datasets have been proposed for different tasks, such as question answering (Liu et al., 2019; Clark et al., 2020; Longpre et al., 2021), natural language inference (Conneau et al., 2018), text summarization (Giannakopoulos et al., 2015; Ladhak et al., 2020; Scialom et al., 2020), numerical reasoning (Shi et al., 2023), code generation (Peng et al., 2024), text-to-SQL (Dou et al., 2023), and readability (Trokhymovych et al., 2024; Naous et al., 2024), among others. Additionally, numerous multilingual datasets have been collected for different tasks (Hu et al., 2020; Ruder et al., 2021; Zhang et al., 2024b; Singh et al., 2024). However, to date, there is no multilingual TATQA dataset, resulting in a lack of evaluation and analysis of multilingual TATQA capabilities and a gap with real scenarios. Therefore, we introduce MULTITAT, a multilingual TATQA dataset, and provide a detailed analysis of the challenges in multilingual TATQA.

5.2 QA Datasets for the Table and Text

Currently, QA datasets for the table and text primarily focus on a single language. For instance, HybridQA (Chen et al., 2020) collects English tables and associated text from Wikipedia. TAT-QA (Zhu et al., 2021), FinQA (Chen et al., 2021), DOCMATH-EVAL (Zhao et al., 2024b), and FinanceMATH (Zhao et al., 2024a) focus on numerical computation in the financial domain, and SciTAT (Zhang et al., 2024a) addresses questions based on tables and text from English scientific papers. However, single-language datasets cannot evaluate the multilingual TATQA capabilities, and overlook the diverse languages in real scenarios. So we propose MULTITAT: the first multilingual TATQA dataset, involving 11 languages and 8 language families. A comparison of MULTITAT and prior works is presented in Appendix A.

The current works on enhancing TATQA performance primarily focus on retrieving relevant information from the context (Luo et al., 2023; Bardhan et al., 2024; Glenn et al., 2024) and generating programs, equations, or step-by-step reasoning process to derive the final answer (Tonglet et al., 2023; Zhu et al., 2024; Fatemi and Hu, 2024). For example, S3HQA (Lei et al., 2023) emphasizes retrieving, where a retriever is initially trained, followed by further filtering based on the question type. Hpropro (Shi et al., 2024) focuses on generating, providing LLMs with commonly used functions to facilitate direct invocation during code generation. However, previous methods are designed for single-language scenarios, directly used to other languages could lead to performance degradation. To address this, we propose OURS, a multilingual baseline that aligns the English TATQA capabilities to other languages.

6 Conclusion

To address the limitations of the existing QA datasets on the hybrid context of tabular and text data (TATQA), we introduce the first multilingual TAT-QA dataset MULTITAT. Specifically, we sample data from mainstream TAT-QA datasets, and translate it into 10 diverse languages. To enhance the TATQA performance in non-English languages, we propose a baseline (OURS). OURS links the relevant information from the hybrid context and reasons in English. We conduct a series of experiments and observe a 19.4% performance drop for non-English languages compared to English. Error analysis reveals that this decline is primarily due to the increased difficulty in linking relevant information in non-English texts and the reduced ability to apply formulas and follow the instructions. Furthermore, OURS achieves an average improvement of 3.3 over other baselines, demonstrating its effectiveness. Analysis suggests that the performance of TATQA across languages is influenced not only by high-resource versus low-resource languages but also by the inherent characteristics of the language.

Limitations

(*i*) MULTITAT only includes single-turn dialogues, leaving multilingual multi-turn dialogues for future work. (*ii*) MULTITAT covers only 11 languages. Future versions should include more languages.

Ethics Statement

All datasets and models used in this paper are publicly available, and our utilization of them strictly complies with their respective licenses and terms of use. Additionally, we confirm that the compensation provided to annotators is significantly higher than the local minimum wage.

References

Elena Angulo, Christophe Diagne, Liliana Ballesteros-Mejia, Tasnime Adamjy, Danish A Ahmed, Evgeny Akulov, Achyut K Banerjee, César Capinha, Cheikh AKM Dia, Gauthier Dobigny, et al. 2021. Non-english languages enrich scientific knowledge: The example of economic costs of biological invasions. *Science of the Total Environment*, 775:144441.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

Jayetri Bardhan, Bushi Xiao, and Daisy Zhe Wang. 2024. Ttqa-rs- a break-down prompting approach for multi-hop table-text question answering with reasoning and summarization. *Preprint*, arXiv:2406.14732.

Anne Beyer, Kranti Chalamalasetti, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2024. clembench-2024: A challenging, dynamic, complementary, multilingual benchmark and underlying flexible framework for llms as multi-action agents. *arXiv preprint arXiv:2405.20859*.

Mahathi Bhagavatula, Santosh GSK, and Vasudeva Varma. 2012. Language independent named entity identification using Wikipedia. In *Proceedings of the First Workshop on Multilingual Modeling*, pages 11–17, Jeju, Republic of Korea. Association for Computational Linguistics.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2024. The impact of language on arithmetic proficiency: A multilingual investigation with cross-agent checking computation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 631–637, Mexico City, Mexico. Association for Computational Linguistics.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Longxu Dou, Yan Gao, Mingyang Pan, Dingzirui Wang, Wanxiang Che, Dechen Zhan, and Jian-Guang Lou. 2023. Multispider: towards benchmarking multilingual text-to-sql semantic parsing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sorouralsadat Fatemi and Yuheng Hu. 2024. Enhancing financial question answering with a multi-agent reflection framework. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 530–537.

Yue Feng, Zhen Han, Mingming Sun, and Ping Li. 2022. Multi-hop open-domain question answering

over structured and unstructured knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 151–156, Seattle, United States. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Parker Glenn, Parag Dakle, Liang Wang, and Preethi Raghavan. 2024. BlendSQL: A scalable dialect for unifying hybrid question answering in relational algebra. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 453–466, Bangkok, Thailand. Association for Computational Linguistics.

Serhii Hamotskyi, Nata Kozaeva, and Christian Hänig. 2024. FinCorpus-DE10k: A corpus for the German financial domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7277–7285, Torino, Italia. ELRA and ICCL.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Yuchen Hu, Chen Chen, Chao-Han Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and EngSiong Chng. 2024. GenTranslate: Large language models are generative multilingual speech and machine translators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 74–90, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Fangyu Lei, Shizhu He, Xiang Li, Jun Zhao, and Kang Liu. 2022. Answering numerical reasoning questions in table-text hybrid contents with graph-based encoder and tree-based decoder. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1379–1390, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Fangyu Lei, Xiang Li, Yifan Wei, Shizhu He, Yiming Huang, Jun Zhao, and Kang Liu. 2023. S3HQA: A three-stage approach for multi-hop text-table hybrid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1731–1740, Toronto, Canada. Association for Computational Linguistics.

Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. 2024. Eliciting better multilingual structured reasoning from LLMs through code. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5169, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Li, Yawei Sun, and Gong Cheng. 2021. Tsqa: tabular scenario based question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13297–13305.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9215–9235, Bangkok, Thailand. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Tongxu Luo, Fangyu Lei, Jiahe Lei, Weihao Liu, Shihu He, Jun Zhao, and Kang Liu. 2023. Hrot: Hybrid prompt strategy and retrieval of thought for table-text hybrid question answering. *Preprint*, arXiv:2309.12669.

Qingkai Min, Yuefeng Shi, and Yue Zhang. 2019. A pilot study for Chinese SQL semantic parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3652–3658, Hong Kong, China. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8383–8394, Torino, Italia. ELRA and ICCL.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. 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*.

Qi Shi, Han Cui, Haofeng Wang, Qingfu Zhu, Wanxiang Che, and Ting Liu. 2024. Exploring hybrid question answering via program-based prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11035–11046, Bangkok, Thailand. Association for Computational Linguistics.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Jonathan Tonglet, Manon Reusens, Philipp Borchert, and Bart Baesens. 2023. SEER : A knapsack approach to exemplar selection for in-context HybridQA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13569–13583, Singapore. Association for Computational Linguistics.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An open multilingual system for scoring readability of Wikipedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6296–6311, Bangkok, Thailand. Association for Computational Linguistics.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085, Online. Association for Computational Linguistics.

Dingzirui Wang, Longxu Dou, and Wanxiang Che. 2022. A survey on table-and-text hybridqa: Concepts, methods, challenges and future directions. *arXiv preprint arXiv:2212.13465*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *Preprint*, arXiv:2407.03658.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xuanliang Zhang, Dingzirui Wang, Baoxin Wang, Longxu Dou, Xinyuan Lu, Keyan Xu, Dayong Wu, Qingfu Zhu, and Wanxiang Che. 2024a. Scitat: A question answering benchmark for scientific tables and text covering diverse reasoning types. *Preprint*, arXiv:2412.11757.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. 2024b. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *Preprint*, arXiv:2411.09116.

Yusen Zhang, Jun Wang, Zhiguo Wang, and Rui Zhang. 2023. XSemPLR: Cross-lingual semantic parsing in multiple natural languages and meaning representations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15918–15947, Toronto, Canada. Association for Computational Linguistics.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024a. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024b. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 310–318, New York, NY, USA. Association for Computing Machinery.

A Comparison with Previous Datasets

In this section, we make a detailed comparison between MULTITAT and previous TATQA datasets, as shown in Table 4. It can be seen that MULTITAT is the first multilingual TATQA dataset, and it gathers previous datasets from three mainstream fields.

We selected these three datasets for the following reasons: (*i*) HybridQA (Chen et al., 2020), the first proposed TATQA dataset, covers multiple topics from Wikipedia and effectively evaluates the ability to locate question-relevant information within hybrid contexts. (*ii*) TAT-QA (Zhu et al., 2021), a mainstream TATQA dataset in the financial domain, requires models not only to identify relevant information but also to possess financial domain knowledge for complex calculations. (*iii*) SciTAT (Zhang et al., 2024a), a recently introduced TATQA dataset in the scientific domain, features nested tables and encompasses a wider variety of reasoning types, including Look Up, Numerical Reasoning, Data Analysis, and Tabulation. To ensure that MULTITAT covers diverse domains and reasoning types as comprehensively as possible, we chose these three datasets.

B Manual Annotation Process

B.1 Annotator Training Process

We hire graduate students majoring in Computer Science who are willing to participate in the annotation process. First, we provide annotators with a clear definition of the task, the specific checks and revisions required (as described in Section §2.2 and §2.3), and instructions on how to use the annotation interface. The annotation interface is shown in §B.2. We also inform them of the annotation deadline and encourage them to discuss any uncertainties with us promptly. Finally, a total of five annotators complete the annotation for §2.2 and §2.3, with a combined time of one month.

B.2 Annotation Interface

In this subsection, we show the interfaces annotated by the annotator, which are developed by ourselves, as shown in Figure 8 and Figure 9.

C Automatic Translation Evaluation

In this section, we present our automated evaluation method for the translation of our dataset, following the prior work (Kocmi and Federmann, 2023). Specifically, for each non-English instance in MULTITAT, we input the table, text, and question, along with their corresponding English counterparts, into Llama3.1-70B (Dubey et al., 2024). We employ the prompt provided by Kocmi and Federmann (2023). to score the translation quality on a scale of 0-100. The scores for each non-English instance are summed and then averaged, with the results presented in Table 5.

D Prompt

In this section, we show the prompts we use to conduct experiments. Table 6 and Table 7 show the prompts of the baselines and OURS in experiments respectively, with French as the example language. The prompt of Three-Agent (Fatemi and Hu, 2024) follows the prompt provided in the original paper. We maintain the unity of demonstrations between different languages and baselines, as shown in Table 7.

E Additional Experiments

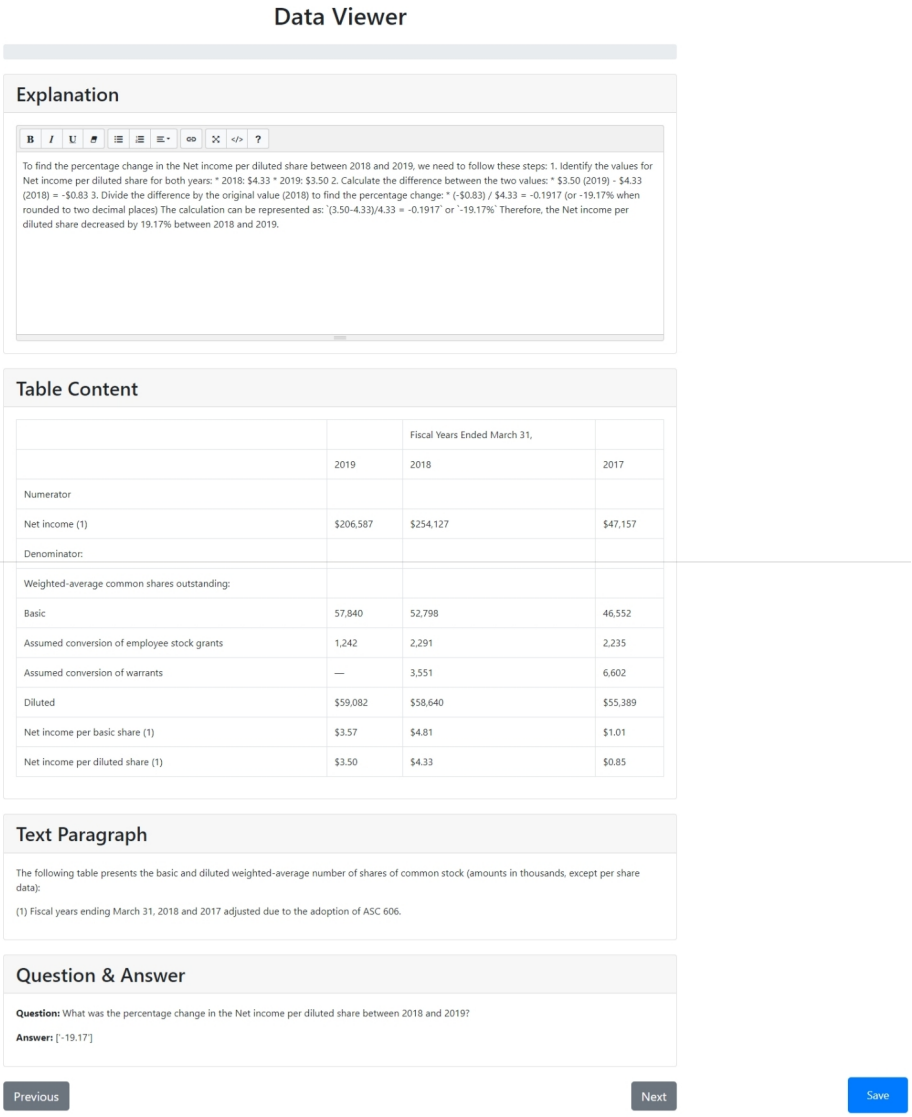
E.1 Other Models

We present the evaluation results of Qwen2.5-Instruct-7B (Yang et al., 2024) on MULTITAT, as detailed in Table 8. Qwen2.5-Instruct, a notable open-source model, exhibits superior performance in code and mathematics and supports over 29 languages (Yang et al., 2024). The results indicate that OURS consistently and significantly outperforms other baselines.

E.2 Other Baselines

In this subsection, we show the results of directly answering the questions (Direct), solving the question with English CoT (Trans-CoT) and PoT (Trans-PoT) after translating the question and context (including the table and text) to English, as shown in Table 9. OURS consistently and significantly outperforms all baseline methods, demonstrating its effectiveness.

Additionally, we observe the following: (*i*) Compared to direct question answering, the overall performance of Native-CoT, Native-PoT, En-CoT, and En-PoT shows substantial improvement (see Table 2). (*i*) The performance of Trans-CoT and Trans-PoT is unstable, primarily due to limitations in the quality of Google Translation. On the one hand, Google Translation struggles to maintain table formatting during translation, especially for low-resource languages such as Bengali and



Data Viewer

Explanation

To find the percentage change in the Net income per diluted share between 2018 and 2019, we need to follow these steps: 1. Identify the values for Net income per diluted share for both years: * 2018: $4.33 * 2019: $3.50 2. Calculate the difference between the two values: * $3.50 (2019) - $4.33 (2018) = -$0.83 3. Divide the difference by the original value (2018) to find the percentage change: * (-$0.83) / $4.33 = -0.1917 (or -19.17% when rounded to two decimal places) The calculation can be represented as: `(3.50-4.33)/4.33 = -0.1917` or `-19.17%` Therefore, the Net income per diluted share decreased by 19.17% between 2018 and 2019.

Table Content

		Fiscal Years Ended March 31,	
	2019	2018	2017
Numerator			
Net income (1)	$206,587	$254,127	$47,157
Denominator:			
Weighted-average common shares outstanding:			
Basic	57,840	52,798	46,552
Assumed conversion of employee stock grants	1,242	2,291	2,235
Assumed conversion of warrants	—	3,551	6,602
Diluted	$59,082	$58,640	$55,389
Net income per basic share (1)	$3.57	$4.81	$1.01
Net income per diluted share (1)	$3.50	$4.33	$0.85

Text Paragraph

The following table presents the basic and diluted weighted-average number of shares of common stock (amounts in thousands, except per share data):

(1) Fiscal years ending March 31, 2018 and 2017 adjusted due to the adoption of ASC 606.

Question & Answer

Question: What was the percentage change in the Net income per diluted share between 2018 and 2019?

Answer: [`-19.17`]

Previous Next Save

Figure 8: The annotation interface is provided to annotators to check the accuracy of the generated rationales.

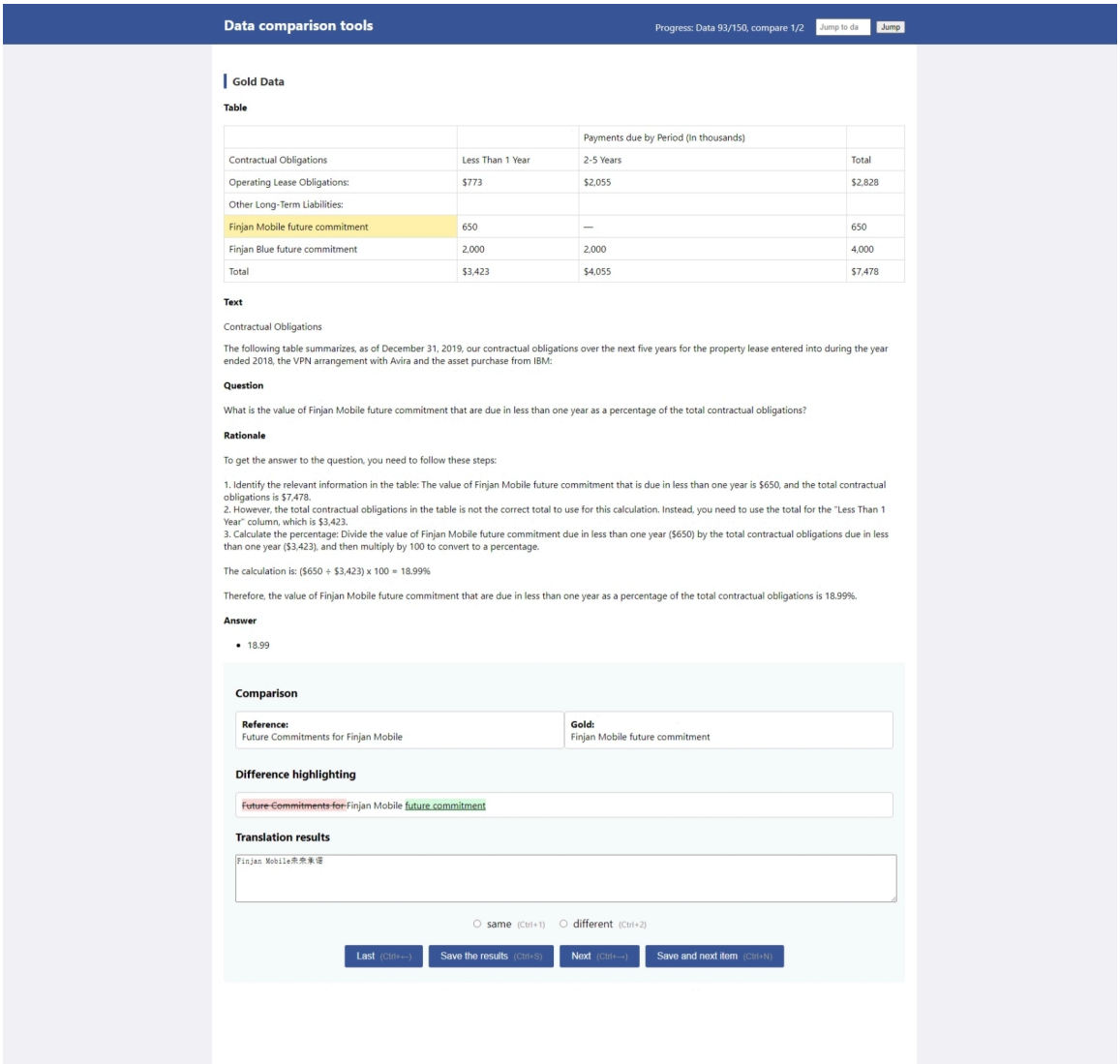


Figure 9: The annotation interface is provided to annotators to check the consistency of the back translation and the original English instance and refine the translated instances.

Dataset	Domain	Language
GeoTSQA (Li et al., 2021)	Geography	Chinese
HybridQA (Chen et al., 2020)	Wikipedia	English
TAT-QA (Zhu et al., 2021)	Finance	English
FinQA (Chen et al., 2021)	Finance	English
QRData (Liu et al., 2024)	Cross	English
DocMath-Eval (Zhao et al., 2024b)	Finance	English
FinanceMATH (Zhao et al., 2024a)	Finance	English
SciTAT (Zhang et al., 2024a)	Science	English
MULTITAT	Wikipedia + Finance + Science	Multilingual

Table 4: Comparison of MULTITAT to previous TATQA datasets.

bn	de	es	fr	ja	ru	sw	te	th	zh	Avg.
94.9	97.1	97.3	97.2	95.8	95.0	94.7	95.9	94.5	94.0	95.6

Table 5: The average translation score of non-English scores in MULTITAT. Avg. denotes the average score of all non-English languages.

Swahili, leading to information loss (Dou et al., 2023). On the other hand, when utilizing back-translation via Google Translation, token consistency with the original table or text cannot be guaranteed.

E.3 Answer Sources

In this subsection, we present the performance of different models and baselines on various answer sources in our dataset, as illustrated in Figure 10 and Figure 11. From Figure 10, it can be observed that multilingual models with better overall performance tend to exhibit smaller performance gaps across different languages. However, even `gpt-4o` still cannot entirely eliminate the discrepancies. From Figure 11, in comparison with Figure 4, OURS demonstrates performance improvements across all answer sources, with a particularly significant enhancement for hybrid answer sources. This is attributed to the ability to better establish connections to relevant information of OURS, thereby mitigating the challenges posed by the heterogeneity of answer sources.

E.4 Answer Types

In this subsection, we present the performance of different models and baselines across various answer types in MULTITAT, as illustrated in Figure 12 and Figure 13. As shown in Figure 12, even for `gpt-4o`, the performance for high-resource languages is consistently superior to that for low-resource languages across different answer types. Figure 13 demonstrates that, compared to Figure 5, OURS reduces the performance gap between languages of varying resource levels to some extent and uniformly improves performance across different answer types.

E.5 Case Study

In this subsection, we show the cases of error types corresponding to the analysis in §4.4, as shown in Figure 14, Figure 15, and Figure 16.

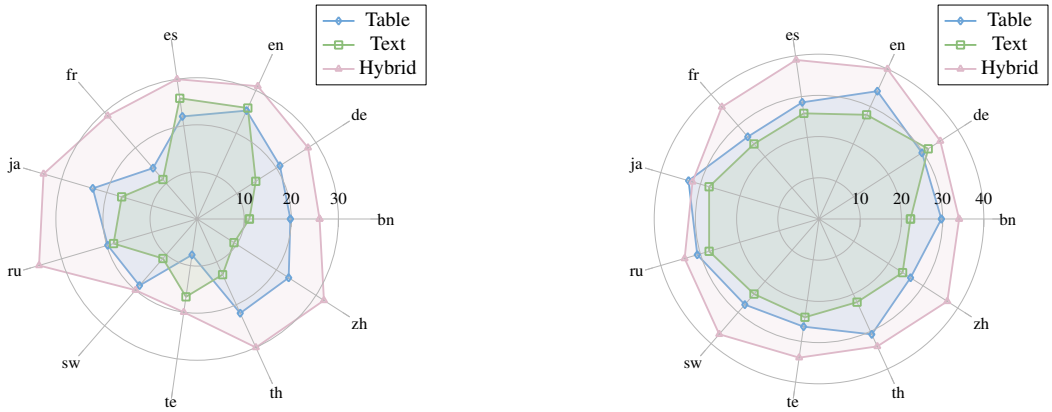


Figure 10: The left part is the EM of OURS across different answer sources on MULTITAT using Llama3.1-8B. The right part is the EM of OURS across different answer sources on MULTITAT using `gpt-4o`.

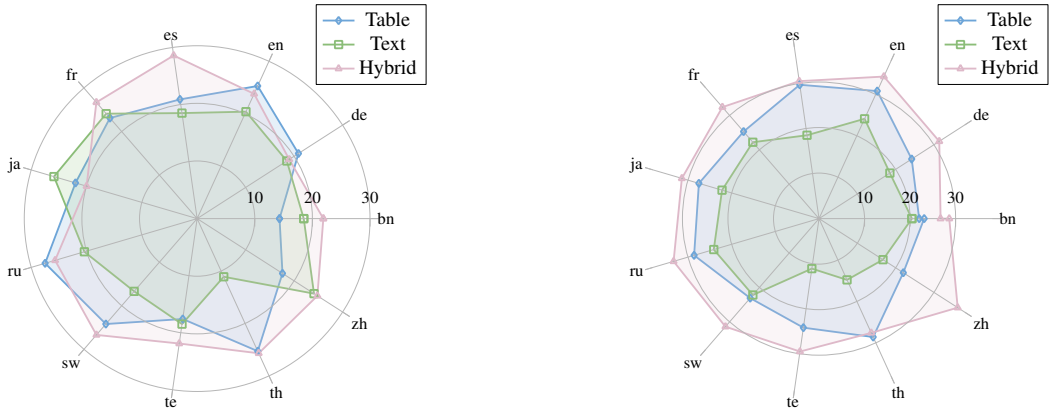


Figure 11: The left part is the EM of En-CoT across different answer sources on MULTITAT using Llama3.1-70B. The right part is the EM of En-PoT across different answer sources on MULTITAT using Llama3.1-70B.

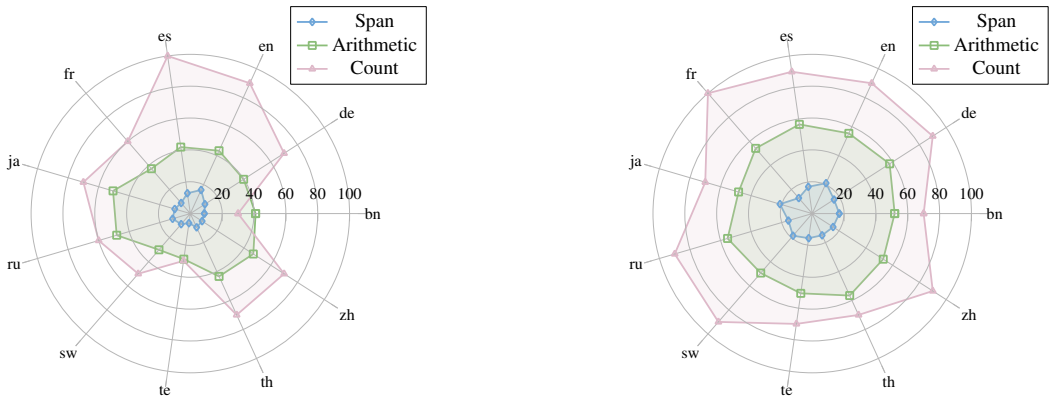


Figure 12: The left part is the EM of OURS across different answer types on MULTITAT using Llama3.1-8B. The right part is the EM of OURS across different answer types on MULTITAT using `gpt-4o`.

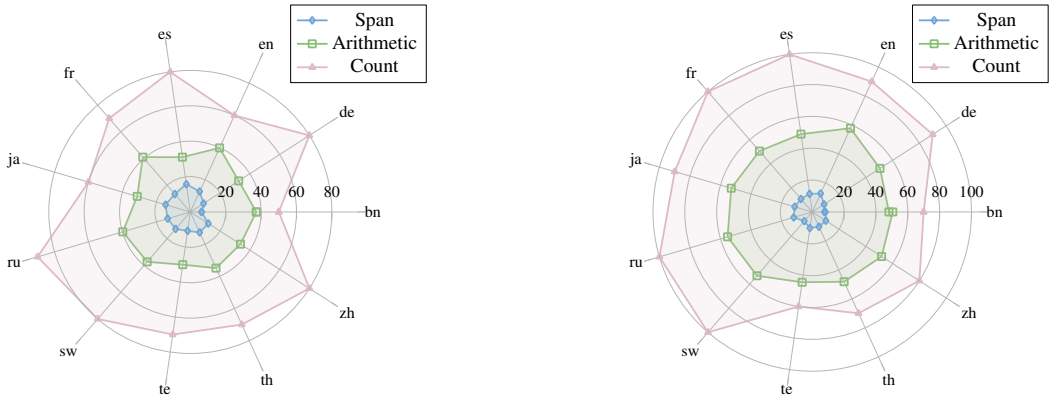


Figure 13: The left part is the EM of En-CoT across different answer types on MULTITAT using Llama3.1-70B. The right part is the EM of En-PoT across different answer types on MULTITAT using Llama3.1-70B.

The prompt for Native-CoT
Lisez le texte et le tableau suivants, puis répondez à une question Voici plusieurs exemples : — {Demonstrations} — Sur la base des exemples ci-dessus, répondez à la question suivante. Représentez votre réponse par : "Explication : <votre explication> Réponse : <votre réponse>" {Table} {Paragraph} Question :{Question}
The prompt for En-CoT
Read the following text and table, and then answer a question. Here are several examples: — {Demonstrations} — Based on the examples above, answer the following question. Represent your answer with: "Explanation: <your explanation> Answer: <your answer>" {Table} {Paragraph} Question :{Question}
The prompt for Native-PoT
Lisez le texte et le tableau suivants, puis écrivez un code Python pour répondre à une question Voici plusieurs exemples : — {Demonstrations} — Sur la base des exemples ci-dessus, répondez à la question suivante avec un code Python. Représentez votre réponse par : "and = <votre réponse>" {Table} {Paragraph} Question :{Question}
The prompt for En-PoT
Read the following text and table, and then write a python code to answer a question Here are several examples: — {Demonstrations} — Based on the examples above, answer the following question with a Python code. Represent your answer with: "ans = <your answer>" {Table} {Paragraph} Question :{Question}

Table 6: The prompts of baselines for French.

The prompt for OURS

Please think in English and locate the relevant information from the text and table according to the question.
Here are several examples:
—
7. Nombre et coûts des employés...
—	2019	2018
——————-	———	———
—	Nombre	Nombre
...
Question: Quelles sont les catégories d'employés listées dans le tableau ?
"Catégories des employés" links to the rows of the table "Opérations clients", "Produit et technologie", "Corporate" and the columns of the table "2019", "2018".
—
Le tableau suivant présente la répartition des revenus par catégorie et segment. ...
Année se terminant le 31 décembre,		
——————————————-	————	————-
	2019	2018
...
Question: En 2019, combien de régions géographiques ont des revenus totaux supérieurs à 20 000 milliers de dollars
"2019" links to the column of the table "2019". "total revenues of geographic regions" links to the rows of the table "Total des revenus de l'Asie-Pacifique", "Total des revenus en Europe", "Total des revenus en Amérique du Nord".
—
Taux d'imposition effectif...
| — | 31 décembre 2019 | 31 décembre 2018 |
...
Question: Quel a été le pourcentage de variation des pertes avant impôts en 2019 ?
"pérdidas antes de impuestos de 2019" y "pérdidas antes de impuestos de 2018" se vinculan a la parte del texto "In 2019 and 2018 we had pre-tax losses of $19,573 and $25,403, respectively".
—
Based on the examples above, analyze the question.
Please note that you **only** need to locate the relevant information, without performing additional calculations.
{Table}
{Paragraph}
Question :{Question}

According to the relevant information, you should also think in English and write a python code to answer the question.
Here are several examples:
—
...
```python
ans = ['Opérations clients', 'Produit et technologie', 'Corporate']
```
—
...
```python
total_revenues_in_all_regions = {'Asie-Pacifique': 6490, 'Europe': 36898, 'Amérique du Nord': 68024}
regions_have_more_than_20000_thousand_total_revenues = [k for k, v in total_revenues_in_all_regions.items() if v > 20000]
ans = len(regions_have_more_than_20000_thousand_total_revenues)
```
—
...
```python
pre_tax_losses_2018 = 25403 pre_tax_losses_2019 = 19573
net_change = pre_tax_losses_2019 - pre_tax_losses_2018
ans = net_change / pre_tax_losses_2018 * 100
```
—
Based on the examples above, answer the question with a Python code.
Please note:
1. In addition to numbers, try to use fr as the answer.
2. Keep your answer **short** with fewer statements.
3. Note the possible minus sign.
4. You MUST generate a Python code instead of returning the answer directly.
Represent your answer with: "ans = <your answer>"
{Table}
{Paragraph}
Question :{Question}

Table 7: The prompts of OURS for French.

Method	bn	de	en	es	fr	ja	ru	sw	te	th	zh	Avg.
Native-CoT	8.4	11.2	16.0	14.4	13.6	15.6	15.6	5.2	6.0	11.6	16.0	12.1
En-CoT	7.2	10.8	14.8	12.0	6.4	12.0	10.0	5.6	8.8	9.6	14.4	10.1
Native-PoT	14.4	13.2	9.6	23.6	21.2	17.2	23.2	10.8	13.3	19.6	20.8	17.0
En-PoT	11.6	14.0	14.0	18.4	12.8	12.8	14.4	4.8	10.4	19.6	10.6	12.2
Three-Agent	12.0	16.8	19.2	16.8	15.2	9.6	16.8	5.6	9.6	13.6	24.4	13.6
OURS	**20.0**	**25.2**	**25.2**	**29.6**	**28.4**	**22.8**	**26.8**	**26.4**	**14.8**	**19.6**	**24.4**	**24.1**
Method	**bn**	**de**	**en**	**es**	**fr**	**ja**	**ru**	**sw**	**te**	**th**	**zh**	**Avg.**
Native-CoT	11.9	17.4	22.1	21.4	19.4	18.0	22.2	9.1	11.2	15.9	25.8	17.7
En-CoT	10.2	14.6	21.2	16.8	9.9	14.5	13.2	8.2	11.7	14.5	25.5	14.6
Native-PoT	15.2	14.1	10.7	24.5	22.9	17.2	24.9	12.1	15.2	20.9	22.0	18.2
En-PoT	12.1	14.6	14.6	18.9	19.3	13.2	15.2	5.7	11.2	20.9	10.7	12.9
Three-Agent	13.9	20.0	26.6	21.2	19.0	11.6	21.0	9.4	11.9	15.2	26.1	18.0
OURS	**21.7**	**26.6**	**26.6**	**32.9**	**30.2**	**24.8**	**28.7**	**28.1**	**15.8**	**21.1**	**26.1**	**26.2**

Table 8: EM/F1 of different models and baselines across languages on MULTITAT using Qwen2.5-Instruct-7B. The best results of each model under each language are annotated in **bold**.

Model	Method	bn	de	en	es	fr	ja
Llama3.1-8b	Direct	10.4/14.0	12.8/17.7	14.8/21.6	13.6/21.1	11.6/17.3	10.4/12.3
	Trans-CoT	2.0/2.4	15.2/16.0	20.8/23.7	18.8/20.6	13.2/13.5	9.6/11.1
	Trans-PoT	2.4/2.5	20.4/21.2	23.2/24.4	21.2/21.6	18.4/18.8	10.8/11.1
	OURS	**20.0/21.3**	**22.4/24.2**	**27.6/31.9**	**25.6/27.8**	**20.0/22.4**	**25.6/26.1**
Llama3.1-70b	Direct	12.4/17.4	21.2/24.5	22.0/26.6	21.6/27.4	18.0/22.3	21.6/24.2
	Trans-CoT	4.4/4.9	20.4/22.0	25.6/29.3	25.6/29.0	16.4/18.1	14.0/14.7
	Trans-PoT	3.2/3.4	22.8/23.8	30.4/32.9	28.4/25.8	22.8/23.7	14.4/14.7
	OURS	**24.0/26.3**	**28.0/31.3**	**31.2/35.3**	**29.4/34.6**	**26.8/31.1**	**26.8/29.4**
Model	**Method**	**ru**	**sw**	**te**	**th**	**zh**	**Avg.**
Llama3.1-8b	Direct	10.8/14.5	9.6/14.7	10.0/13.7	12.0/14.1	11.2/19.3	11.6/16.4
	Trans-CoT	16.0/18.0	9.2/10.2	9.2/9.6	11.6/13.1	4.8/8.4	11.9/13.3
	Trans-PoT	21.2/22.5	16.0/16.4	14.8/15.1	13.6/14.9	6.4/7.2	15.3/15.8
	OURS	**25.2/27.0**	**17.2/20.0**	**14.4/15.2**	**22.8/24.6**	**23.6/28.0**	**22.2/24.4**
Llama3.1-70b	Direct	20.4/23.4	20.0/23.3	16.8/20.1	20.4/23.5	19.6/28.5	19.5/23.7
	Trans-CoT	21.2/22.9	17.6/19.6	14.8/16.4	19.6/21.9	9.2/12.9	17.0/18.4
	Trans-PoT	24.0/24.8	20.0/20.9	19.6/20.5	18.0/19.5	9.6/12.4	19.6/20.4
	OURS	**28.8/33.5**	**30.8/34.7**	**22.8/25.9**	**26.8/30.5**	**28.0/34.9**	**27.6/31.6**

Table 9: EM/F1 of different models and baselines across languages on MULTITAT. The best results of each model under each language are annotated in **bold**.

Text

The Nigerian Television Authority or **NTA** … is widely viewed as the authentic voice of the Nigerian government .

Table

Channel	**Label**	Format	Programming
		…	
15.7	**KVVV 15**	480i [0]	**NTA** [3]

Question

What is the **label** of the programming that is widely considered as the reliable voice of the Nigerian government ?

Linking 🇬🇧 English

The relevant information is located in the table, specifically in the row where the "Programming" column mentions "NTA [3]". The corresponding "Label" is "KVVV 15".

Reasoning

ans = 'KVVV 15'

Linking 🇩🇪 German

Die Kennung des Programms, das weithin als die verlässliche Stimme der nigerianischen Regierung angesehen wird, ist NTA.

Reasoning

ans = 'NTA'

Figure 14: The case for the error type of "Linking".

Text

\paragraph {Efficiency.} To evaluate the efficiency, we present in Table~\\ref{tab:runtime} the averaged time needed to predict one token for different methods ...

Table

Method	Time per Token
…	
Ensemble Top-3 ($\rho_1=0.4, \rho_{10}=0$)	0.2085 s

Question

Given that the ensemble top-3 method is used in a scenario where the source sentence is revealed at a speed of 0.6 seconds per token (as in general speech), what is the maximum percentage of the source sentence that can be processed in real-time using the ensemble method, assuming no additional time for processing or transmission delays?

Linking 🇬🇧 English

"ensemble top-3 method" links to the row of the table "Ensemble Top-3 ($\rho_1=0.4, \rho_{10}=0$)". … links to the column of the table "Time per Token".

Reasoning

ans = **0.2085 / 0.6 * 100**

Linking 🇧🇩 Bengali

প্রতি টোকেন সময় হল ০.২০৮৫ সেকেন্ড। সুতরাং, প্রতি সেকেন্ডে প্রক্রিয়াকরণ করা টোকেনের সংখ্যা হল ১/০.২০৮৫ = ৪.৭৯৭। … **৪.৭৯৭/১.৬৬৭ = ২.৮৮ বা ২৮৮%।**

Reasoning

ans = 288

Figure 15: The case for the error type of "Formula".

Text

Marine Services Segment GMSL generally generates revenue by providing maintenance services for subsea telecommunications cabling ...

Table

Years Ended December 31,	2019	2018
Telecommunication - Maintenance	$86.8	$87.0
	…	

Question

What is the increase / (decrease) in the telecommunication maintenance from 2018 to 2019?

Linking 🇬🇧 English

"2018 telecommunication maintenance" and "2019 telecommunication maintenance" link to the rows "Telecommunication - Maintenance" in the columns "2018" and "2019".

Reasoning

telecommunication_maintenance_2018 = 87.0

telecommunication_maintenance_2019 = 86.8

ans = telecommunication_maintenance_2019 - telecommunication_maintenance_2018

Linking 🇨🇳 Chinese

根据表格中的数据，2018年电信维护的收入为$87.0，2019年电信维护的收入为$86.8。因此，2018年和2019年电信维护减少了$0.2。

Reasoning

telecom_maintenance_2018 = 87.0 telecom_maintenance_2019 = 86.8

net_change = telecom_maintenance_2019 - telecom_maintenance_2018

ans = f"**减少了{abs(net_change)}百万美元**"

Figure 16: The case for the error type of "Redundancy".