

1+1>2: Can Large Language Models Serve as Cross-Lingual Knowledge Aggregators?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have garnered significant attention due to their remarkable ability to process information across various languages. Despite their capabilities, they exhibit inconsistencies in handling identical queries in different languages, presenting challenges for further advancement. This paper introduces a method to enhance the multilingual performance of LLMs by aggregating knowledge from diverse languages. This approach incorporates a low-resource knowledge detector specific to a language, a strategic language selection process, and mechanisms for answer replacement and integration. Our extensive experiments demonstrate notable performance improvements, particularly in reducing language performance disparity. An ablation study confirms that each component of our method significantly contributes to these enhancements. This research highlights the inherent potential of LLMs to harmonize multilingual capabilities and offers valuable insights for further exploration.

1 Introduction

Large Language Models (LLMs) are increasingly recognized for their impressive capabilities in natural language processing (NLP). Employed across a variety of domains such as the medical sector (Liu et al., 2023c; Zhang et al., 2023a), software engineering (Qian et al., 2023), scientific research (Guo et al., 2023; Li et al., 2024c), and LLM-based applications (Liu et al., 2023b; Guo et al., 2024; Huang et al., 2024), LLMs have demonstrated significant utility. Additionally, recent advancements in LLMs have expanded research (Qin et al., 2024; Li et al., 2024a; Xu et al., 2024b; Chen et al., 2024), which focuses on enhancing their ability to process multiple languages and thereby increasing their accessibility and relevance across diverse linguistic demographics.

Despite these advancements, LLMs demonstrate

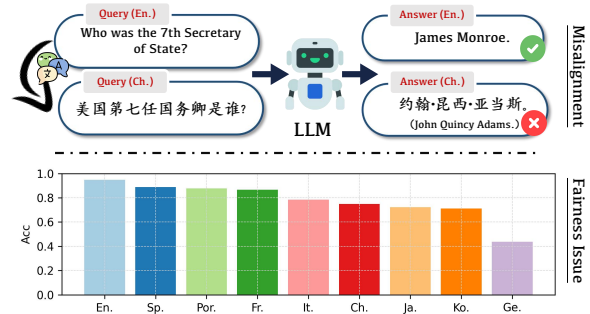


Figure 1: The top is an example of distinct answers to the same questions in different languages. The bottom is the GPT-4’s performance on 300 queries in HalluEval (Li et al., 2023a) of nine different languages.

inconsistencies when processing queries in different languages with the same meaning (Li et al., 2024d), as evidenced by the results in Figure 1. This inconsistency not only diminishes the efficacy and fairness of LLMs but also signals underlying knowledge conflicts (Xu et al., 2024a) that prevent these models from achieving true intelligence (Liu et al., 2023b; Huang et al., 2024). Furthermore, such inconsistency can erode trust in LLM applications, particularly when users from varied linguistic backgrounds cannot equally benefit from the technology (Li et al., 2023b).

To address the inconsistency problems in LLMs, we propose a novel method by leveraging the intrinsic capabilities of LLMs through integrating knowledge across different languages. Our approach begins with the development of a low-resource knowledge detector. This detector assesses whether a user’s query involves knowledge that is underrepresented in the specific language. When the query does not feature low-resource knowledge, it is directly addressed by the LLMs. In contrast, if low-resource knowledge is detected, the LLMs will be required to select the most relevant target language to handle this specific knowledge. Once the target language is selected, the query is translated into

068 this language, and the LLMs generate a response
069 based on the translated query. This response either
070 replaces the original answer or is integrated with
071 it. Finally, the response is translated back to the
072 original language of the query and delivered to the
073 user.

074 We conducted comprehensive experiments us-
075 ing six popular LLMs and five bilingual datasets
076 (specifically, English and Chinese) to evaluate the
077 effectiveness of our proposed method. The ex-
078 perimental results demonstrate that our approach
079 successfully integrates knowledge from different
080 languages, leading to the improvement in over-
081 all performance. Importantly, it significantly re-
082 duces the performance disparities across languages,
083 thereby addressing the inconsistency issues inher-
084 ent in LLMs and promoting fairness for down-
085 stream applications. Additionally, our ablation
086 study confirms that both the low-resource knowl-
087 edge detector and the language selection process
088 are crucial to the improvements observed. Overall,
089 our contributions are as follows:

- 090 • We posed an important challenge on the incon-
091 sistency of LLMs in downstream tasks, and the
092 low-resource knowledge in a specific language
093 can be brought from another language.
- 094 • Based on the observation, we propose a method
095 that utilizes the LLMs’ internal capability to en-
096 hance its performance on datasets in different
097 datasets through a low-resources knowledge de-
098 tector, language selection process, and answer
099 replacement & integration.
- 100 • We conduct extensive experiments on six popular
101 LLMs and five bilingual datasets. The results
102 show that our proposed method effectively en-
103 hances the performance of LLMs by integrating
104 knowledge from different languages and reduce
105 the performance gap in different languages.

106 2 Related Work

107 2.1 Multilingual LLMs

108 There has been a surge in research and work on
109 Multilingual Large Language Models (MLLMs)
110 (Qin et al., 2024; Li et al., 2024a; Xu et al., 2024b;
111 Chen et al., 2024; Etxaniz et al., 2023). For in-
112 stance, the InternLM, proposed by Team (2023),
113 is a multilingual language model that has demon-
114 strated excellent performance on multiple Chinese
115 benchmarks. Similarly, PolyLM (Wei et al., 2023b)
116 is another LLM trained using curriculum learning,
117 surpassing other open-source models in multilin-

118 gual tasks. Besides the above multilingual LLMs,
119 the popular LLMs also include the ChatGLM se-
120 ries developed by Du et al. (2022) and Zeng et al.
121 (2022), and Baichuan series Yang et al. (2023). To
122 improve model performance on multilingual tasks,
123 Muennighoff et al. (2023) and Zhang et al. (2023b)
124 focus on utilizing multilingual training data to fine-
125 tune the parameters.

126 In terms of evaluation, Lai et al. (2023) as-
127 sessed ChatGPT’s performance across 37 differ-
128 ent languages. CulturaX (Nguyen et al., 2023) is a
129 multilingual dataset containing 6.3 trillion tokens
130 across 167 languages, aimed at promoting the de-
131 velopment of multilingual LLMs. Additionally,
132 M3Exam (Zhang et al., 2023c) introduces a dataset
133 derived from real and official human exam ques-
134 tions, designed for evaluating LLMs in a multilin-
135 gual, multimodal, and multilevel context. BUFFET
136 consolidates 15 varied tasks across 54 languages
137 into a sequence-to-sequence format, offering a stan-
138 dardized set of few-shot examples and instructions
139 (Asai et al., 2023).

140 2.2 Factuality in LLMs

141 One way to improve the factuality of LLMs is the
142 utilization of knowledge graphs (KGs)(Sun et al.,
143 2024b). For instance, Abu-Rasheed et al. (2024)
144 uses knowledge graphs to learn explainable recom-
145 mendations. Yang et al. (2024b) suggests improv-
146 ing LLMs through the development of knowledge
147 graph-enhanced LLMs, which offers a method to
148 boost the factual reasoning capabilities of LLMs.
149 (Sun et al., 2024a) utilizes the LLM as an agent to
150 interact with and navigate through the KGs, iden-
151 tifying relevant entities and relationships, and con-
152 ducting reasoning with the knowledge it gathers.

153 Another method to enhance the factual knowl-
154 edge of LLMs is the utilization of prompt engi-
155 neering. Previous studies propose various prompt
156 methods such as Chain-of-Thoughts (CoT) (Wei
157 et al., 2023a) and Tree-of-Thoughts (ToT) (Yao
158 et al., 2023). Moreover, some studies use knowl-
159 edge injection to enhance the domain capability of
160 LLMs (Huang and Sun, 2024).

161 2.3 Hallucination Mitigation

162 A significant challenge associated with LLMs is
163 their tendency to generate seemingly plausible yet
164 fabricated responses, a phenomenon known as hal-
165 lucination (Sun et al., 2024c). To address this issue,
166 recent research has introduced various hallucina-
167 tion mitigation strategies (Tonmoy et al., 2024).

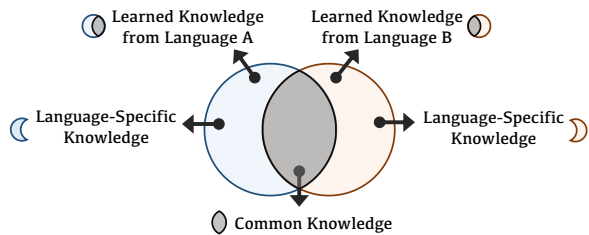


Figure 2: The knowledge domain of a multilingual LLM can be separated into multiple sections (the figure shows two). The language-specific knowledge (pure blue or pure orange) in one language can be utilized for improving the performance in other languages.

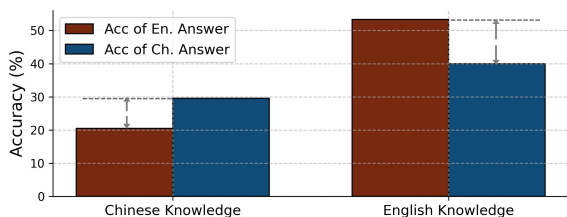


Figure 3: The average performance of six LLMs in five datasets. We show the accuracy of Chinese and English domain knowledge with the query/answer in Chinese and English.

For example, Feng et al. (2024) leverage multi-LLMs collaboration to decrease hallucinations in LLM outputs. Additionally, Guan et al. (2024) have developed a novel framework called Knowledge Graph-based Retrofitting (KGR), which integrates LLMs with KGs to minimize factual hallucinations during reasoning. Similarly, Manakul et al. (2023) propose SelfCheckGPT, a sampling method that verifies the accuracy of responses from black-box models without the need for an external database.

3 Methodology

3.1 Motivation

Our proposed method draws inspiration from the distinct knowledge domains inherent to different languages. As illustrated in Figure 2, language-specific knowledge can serve as supplementary information for another language. Figure 3 demonstrates that when queries related to English domain knowledge are posed in Chinese, the performance (i.e., accuracy) of LLMs declines compared to those posed in the English language. Furthermore, Figure 8 reveals that LLMs often provide correct answers in only one of two languages for a given query, suggesting the potential to use the correct response to rectify inaccuracies in the other language. These observations underscore the po-

tential to leverage the strengths of each language to enhance LLM performance across different languages. As shown in Figure 4, the proposed method includes three main modules: *low-resource knowledge detection*, *target language selection*, and *answer replacement & integration*.

3.2 Construction of Low-Resource Dataset

We first construct a low-resource dataset to measure current LLMs’ multilingual knowledge transfer capacity. We also use this dataset to train our Low-Resource Knowledge Detector in section 3.3. We initial the dataset with the combination of various existing question-answering datasets including TriviaQA (Joshi et al., 2017), CMMLU (Li et al., 2024b), HalluEval (Li et al., 2023a), TruthfulQA (Lin et al., 2022), MKQA (Longpre et al., 2021), XQuAD (Artetxe et al., 2019), LC-QuAD (Trivedi et al., 2017), KgCLUE (Xu et al., 2020). Moreover, we also construct a dataset that uses LLM-powered synthesized data to cover more knowledge and topics in the training corpus (We call it MULTIGEN). The details of the constructed dataset are shown in Appendix C.

To label these data items, we first use an LLM-Human collaboration to label the samples as Chinese-specific, English-specific, or common sense. Specifically, to confirm the correctness of the labeling, we infer the GPT-4 twice to label the samples with a temperature of 1.0 to enlarge the potential uncertainty of its output. We then conduct human inspections of the dataset where the labels are inconsistent in two labeling processes, to confirm the labeling and filter out the samples that are too hard or ambiguous for current LLMs. The statistics of the dataset can be found in Table 1.

3.3 Low-Resource Knowledge Detector

The multilingual misalignment stems from the unbalanced training data as the knowledge with low data resources is less likely to be captured by the language model during the pretraining process. For example, queries about the details of Chinese history are not well answered by the model if asked in English as they appear less frequently in the English pretraining corpus. This phenomenon could be improved by fully utilizing the model’s inherent capacity. To implement this process, we first adopt a low-resource knowledge detector to identify these low-resource queries and later borrow knowledge from other languages for help.

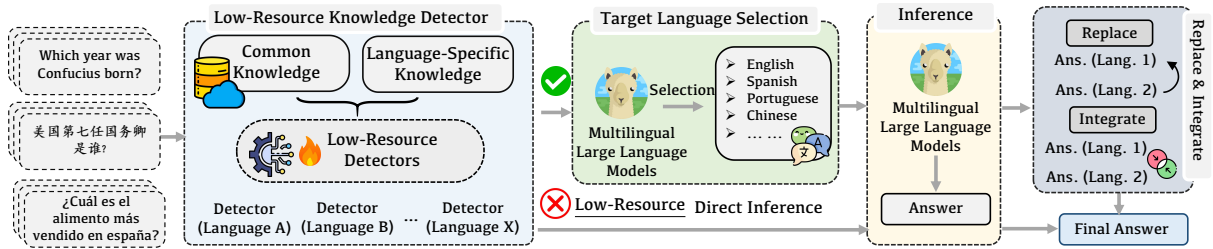


Figure 4: The proposed method begins with the query detection of low-resource knowledge powered by a detector. If low-resource knowledge is detected within the queries, LLMs then select the language most likely to yield the best answer. Answer replacement and integration are employed to formulate the final response.

We train a classifier for each source language to identify the low-resource query for that language. This classifier separates the query about common sense and language-specified knowledge (e.g. Spanish query about Spanish culture) from the low-resources query (e.g. Spanish query about Turkish geography). Queries of the former class are fed into the normal pipeline of language generation while the latter queries are to be enhanced by the knowledge of other languages through our design of other modules. Given a query x in the original language L_o , the low-resource knowledge detector F_{L_o} works as follow:

$$F_{L_o}(x) = \begin{cases} 1 & , \quad x \text{ is low-resource query of } L_o \\ 0 & , \quad \text{else} \end{cases} \quad (1)$$

We demonstrate in the experiment that a classifier is effective enough to distinguish low-resource queries from others. The construction of the training dataset of F_L can be found in subsection 4.1.

The method is cost-effective as it does not require the translation of all queries to multiple languages considering that low-resource query is only a small part of user queries. The majority of user queries are related to common sense and knowledge specified in that language and do not need to go through the following process.

3.4 Target Language Selection

After selecting the low-resource query from the user’s input, we later adopt a target language selection module to find the most suitable language for that question (e.g. translating a question in English about Chinese history into Chinese). Answering the query with its most resourceful language would improve output quality in terms of correctness and may offer more useful details to the user. We implement this process by prompting the LLM itself as the selection is model-dependent. Different LLMs

may select different target languages due to their pretraining corpus. Given the prompt P_{sel} to help select the target language, the low-resource query x , the procedure of Target Language Selection is defined as follows:

$$x' \leftarrow \text{Trans}(x, \text{LLM}(P_{sel}(x))), \quad (2)$$

where translator $\text{Trans}(Q, L_t)$ translates the input Q into target language L_t , and LLM is the large language model that selects the most suitable language for x with prompt P_{sel} .

3.5 Answer Replacement & Integration

After translating the original query x to the query in target language x' , we use it to prompt the model for the answer in target language a_t . We simply translate the answer back to the original language to get the final answer a_{final} for the user’s understanding.

$$a_{final} \leftarrow \text{Trans}(a_t, L_o), \quad (3)$$

where L_o is the original language of the user’s query.

We also explore the integration of answers in the scenario of open-ended question answering (the prompt template is shown in Appendix F). We let the LLM combine and integrate the answer in the target language a_t and the answer in the original language $a_o = \text{LLM}(x)$:

$$a_{final} = \text{LLM}(P_{int}(a_t, a_o)), \quad (4)$$

where P_{int} is the prompt to help LLM integrate between a_t and a_o , and a_{final} is the final answer.

4 Experiments

We chose English and Chinese for our experiments primarily due to their broad applicability and the availability of resources. Firstly, most LLMs,

Algorithm 1 Proposed Method

Require: Query x in original language L_o

Ensure: Final answer a_{final}

- 1: **Low-Resource Knowledge Detection:**
 - 2: Train classifier F_{L_o} for language L_o
 - 3: $isLowResource \leftarrow F_{L_o}(x)$
 - 4: **if** $isLowResource == 1$ **then**
 - 5: **Target Language Selection:**
 - 6: Define prompt P_{sel} for selecting target language
 - 7: $L_t \leftarrow \text{LLM}(P_{\text{sel}}(x))$
 - 8: $x' \leftarrow \text{Trans}(x, L_t)$
 - 9: **Answer Generation:**
 - 10: $a_t \leftarrow \text{LLM}(x')$
 - 11: $a_o \leftarrow \text{Trans}(a_t, L_o)$
 - 12: **Answer Integration:**
 - 13: Define prompt P_{int} for integrating answers
 - 14: $a_{\text{final}} \leftarrow \text{LLM}(P_{\text{int}}(a_t, a_o))$
 - 15: **else**
 - 16: $a_{\text{final}} \leftarrow \text{LLM}(x)$
 - 17: **end if**
 - 18: **return** a_{final}
-

Dataset	Chinese	Common	English	Total	Lang.
TriviaQA	21	754	1040	1815	En.
CMMLU	1200	2162	2751	6113	Ch.
HalluEval	28	923	1033	1984	En.
TruthfulQA	9	322	212	543	En.
MKQA	71	315	1114	1500	En.
XQuAD	72	610	503	1185	En.
LC-QuAD	2	640	345	987	En.
KgCLUE	1218	610	172	2000	Ch.
MULTIGEN	1095	1121	1083	3299	En.
Total	3716	7457	8253	19426	/

Table 1: Dataset statistics of the low-resource knowledge detector. "Lang." is the original language for the dataset.

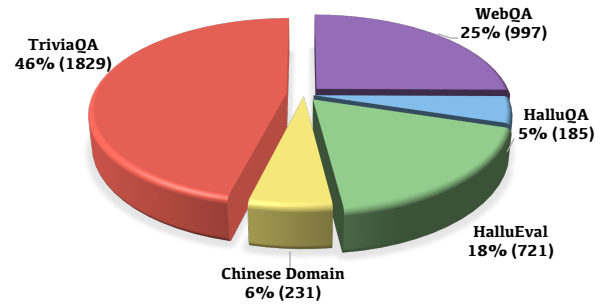


Figure 5: Statistics of the dataset in our experiments.

particularly open-source ones like the ChatGLM series, perform significantly in English and Chinese. This trend highlights the advanced development and optimization of LLMs for these languages, making them ideal for rigorous testing. Secondly, major LLM benchmarks and datasets predominantly focus on these two languages. For instance, besides English benchmarks or datasets, benchmarks such as HalluQA and AlignBench are primarily designed around English and Chinese, providing a robust framework for evaluating our methods. Lastly, the linguistic features and data availability in English and Chinese ensure comprehensive evaluation and validation of our approaches and suggest that our findings could be extrapolated to other languages. This potential for cross-linguistic application supports the broader relevance and utility of our study, choosing English and Chinese as both strategic and impactful.

4.1 Experiment Setup

Training Datasets for Detectors. As we need to train the low-resource detector for each language, for the dataset in English (*e.g.*, TriviaQA) or the dataset in Chinese (*e.g.*, CMMLU, KgCLUE), we translate them to another language (*i.e.*, Chinese or

English) through translation API ¹.

Detailed Setting. To ensure the reproducibility of results, the temperature parameter for all LLMs is set to 0. For ChatGPT, GPT-4, and Qwen-turbo, we use the official API. For Yi-34b, we use the API from Replicate². For ChatGLM3 and Llama3-Chinese, we deploy them locally for inference with a V100 (40G).

Test Datasets. We selected five datasets for our study, comprising four pre-existing datasets and one that we developed in-house. The following criteria guided our selection:

- The datasets should not predominantly consist of common-sense questions (*i.e.*, questions that are independent of linguistic background), as this minimizes the potential for LLMs to demonstrate improvement through linguistic knowledge.
- The datasets should maintain a balance in difficulty; they should not be overly simplistic or excessively challenging. Datasets that are too easy can lead to inflated performance metrics for LLMs, thereby reducing the potential for meaningful improvement. Conversely, datasets that are too challenging can degrade performance across all linguistic contexts, thereby constrain-

¹<https://fanyi.youdao.com/>

²<https://replicate.com/>

ing the opportunity to enhance performance in the target language by leveraging knowledge of additional languages.

For all datasets in our study, we select QA-pair samples from them and do not use other extra data to facilitate our evaluation. Totally, we select five datasets for evaluating our method. These include four existing dataset: TriviaQA (Joshi et al., 2017), HaluEval (Li et al., 2023a), HalluQA (Cheng et al., 2023), and WebQA (Li et al., 2016). We show the statistics of the datasets we selected in Figure 5 and the details are shown in Appendix A. In addition to the four datasets mentioned above, we have constructed a bilingual Chinese-English dataset tailored to the Chinese domain. Details of the construction process are provided in Appendix D.

Models. We carefully select six popular LLMs including proprietary and open-source LLMs that master both English and Chinese: ChatGPT (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), ChatGLM3 (Zeng et al., 2022; Du et al., 2022), Yi-34b (AI et al., 2024), Qwen-turbo (Bai et al., 2023), and Llama3-Chinese (Ila, 2024).

4.2 Main Results

We evaluate the effectiveness of our proposed method on five benchmark datasets and six popular LLMs mentioned above. Each dataset is translated into a Chinese and an English version for later assessment. We first infer the models with the queries in the dataset to get the generated answers. We then leverage GPT-4 as the judge model to compare each generated answer with the reference answer in the dataset to see if the model produces a correct output. We calculate the generation accuracy and present the result in Table 2. We mark the result in green where there is a significant improvement of more than 1% and mark the result in red if the accuracy decrease by more than 1%.

As can be seen from the table, our method can effectively improve the performance of the model in many scenarios. To be specific, the performance of the GPT-4 model on the HalluEval dataset in Chinese improves significantly from 47.99% to 64.36%. This means there still exists a large cross-lingual knowledge gap in advanced models such as GPT-4 and our method successfully leverages the knowledge across languages to enhance the model’s performance. It is important to notice that the improvements do not rely on other models or online resources, they exist due to our leverage of

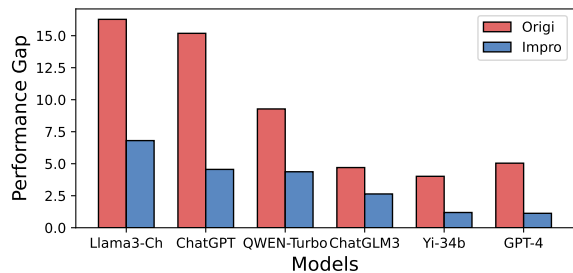


Figure 6: The average performance gap on datasets before and after applying our method.

the model’s inherent capacity.

It can also be observed from Table 2 that most improvements happen in the language that is different from that of the original dataset, which is also the part where the models suffer from a weaker performance. The comparison of the cross-lingual performance gap before and after applying our method is shown in Figure 6. The figure showcases that our method could significantly reduce the knowledge gap between languages in all LLMs we evaluate, thus improving the fairness of the application for users of different languages.

4.3 Ablation Study

As our generation pipeline consists of several parts, we conduct an ablation study to validate their effectiveness and expenses.

The Impact of the Low-resource Detector. The low-resource detector serves as a filter to sift the language-specific queries from the majority of the queries that involve only commonsense, thus improving efficiency and reducing the expense of the pipeline. As can be observed in Figure 7, a low-resource query detector would significantly reduce the average inference time per sample from more than 9 seconds to less than 6.5 seconds if the ratio of the low-resource queries is 0.05 in the dataset. When the ratio of the low-resource query in the dataset increases, the detector passes more samples into the translation pipeline and increases the average inference time.

Another intriguing finding is that the low-resource detector would increase the model performance. As shown in Table 3, the performance of the pipeline is unstable when we remove the low-resource detector. The overall performance would also drop as we observed in Figure 7. This indicates that the detector and LLM itself can be complementary. The full result of the models’ performance without the low-resource detector can be

Dataset	Lang.	ChatGLM3		ChatGPT		GPT-4		Yi-34b		Qwen-turbo		Llama3-Ch.	
		Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.
HalluEval	(en)	18.03%	18.03%	57.98%	57.84%	67.13%	67.13%	42.86%	42.72%	29.31%	29.31%	40.67%	40.67%
	(ch)	11.23%	17.34%	32.07%	51.40%	47.99%	64.36%	25.10%	39.67%	19.35%	26.09%	25.35%	37.19%
HalluQA	(en)	20.00%	25.95%	34.27%	30.90%	51.89%	54.05%	38.38%	47.03%	25.97%	37.57%	22.83%	19.57%
	(ch)	22.16%	22.16%	21.91%	24.16%	49.73%	51.35%	45.95%	44.86%	43.65%	43.09%	15.22%	16.30%
Chinese Domain	(en)	9.52%	20.78%	41.85%	42.73%	56.71%	58.44%	33.33%	55.84%	27.19%	46.05%	30.73%	24.24%
	(ch)	32.47%	32.47%	41.85%	41.85%	59.31%	59.74%	63.64%	63.20%	62.28%	61.84%	18.61%	18.61%
triviaQA	(en)	36.32%	36.32%	90.53%	90.37%	94.09%	94.09%	79.33%	79.17%	59.59%	59.47%	77.27%	77.16%
	(ch)	21.33%	31.95%	54.60%	82.67%	82.77%	91.90%	59.43%	75.56%	41.53%	52.99%	43.92%	65.17%
WebQA	(en)	28.51%	38.15%	59.08%	58.88%	67.70%	69.41%	57.07%	68.71%	49.48%	61.08%	50.00%	48.09%
	(ch)	48.69%	48.49%	57.35%	57.86%	72.52%	72.42%	76.93%	76.13%	71.12%	71.33%	37.02%	38.43%

Table 2: Six LLMs’ performance on our proposed method.

Dataset	Lang.	Yi-34b		Qwen-turbo		Llama3-Ch.	
		Orig.	Impr.	Orig.	Impr.	Orig.	Impr.
HalluEval	(en)	42.86%	41.75%	29.31%	29.59%	40.67%	40.67%
	(ch)	25.10%	39.81%	19.35%	26.51%	25.35%	37.33%
HalluQA	(en)	38.38%	47.03%	25.97%	37.57%	22.83%	18.48%
	(ch)	45.95%	45.95%	43.65%	39.78%	15.22%	20.65%
Chinese Domain	(en)	33.33%	57.58%	27.19%	48.25%	30.74%	24.24%
	(ch)	63.64%	57.14%	62.28%	62.28%	18.61%	22.51%

Table 3: Selected LLMs’ performance on the setting without a low-resource detector.

Dataset	Lang.	Yi-34b		Qwen-turbo		Llama3-Ch.	
		Orig.	Impr.	Orig.	Impr.	Orig.	Impr.
HalluEval	(en)	42.86%	42.72%	29.31%	29.17%	40.67%	40.25%
	(ch)	25.10%	42.58%	19.35%	28.61%	25.34%	39.97%
HalluQA	(en)	38.38%	46.49%	25.97%	38.67%	22.83%	21.74%
	(ch)	45.95%	44.86%	43.65%	41.44%	15.21%	15.76%
Chinese Domain	(en)	33.33%	61.04%	27.19%	58.33%	30.74%	19.48%
	(ch)	63.64%	60.17%	62.28%	58.33%	18.61%	18.61%

Table 4: Selected LLMs’ performance on the setting without language selection.

found in Table 7.

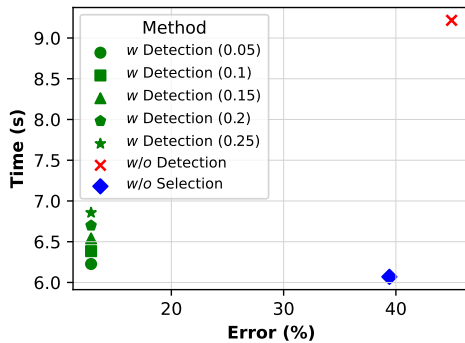


Figure 7: The relationship of time efficiency and error rate. The error rate is the percentage sum of all decreasing in five datasets (value in red on Table 2, Table 3 (w/o Detection), Table 4 (w/o Selection)).

The Impact of the Language Selection Module. The language selection module can choose the proper language to answer the question with model-specific choice. It is also flexible in the multi-lingual setting as the resulting target language can be more than two as we test. However, we still validate its effectiveness in the bi-lingual setting, comparing it with the strategy of using the opposite language when the query is detected as low-resources, and show our result in Table 4. The

trade-off between its cost and error can also be found in Figure 7.

As we can see from Figure 7, the language selection module only adds a small inference cost while significantly improving the model performance. This is due to the existence of the query that is low-resource for both languages, in which switching to the opposite language may make the situation worse. In these situations, the language selection module may pick a third language to better answer the question. The full result of the performance without the language selection module can be found in Table 8.

The Comparison between Answer Replacement and Integration. We further investigated the effectiveness of answer replacement and integration strategies. Given that QA setups with a golden answer may not always accommodate answer integration effectively (for example, when the answers in two different languages factually conflict), we opted for a subset in AlignBench (Liu et al., 2023a) as our evaluation dataset. AlignBench provides a comprehensive, multi-dimensional benchmark designed to assess the alignment of LLMs in Chinese, featuring a variety of open-ended questions. To

Type	Lang.	ChatGPT		GPT-4		ChatGLM3		Yi-34b		Qwen-Turbo		Llama3-Ch.	
		Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.
Integrate	(ch)	4.98	5.26	6.90	6.85	4.15	4.09	5.82	5.91	5.88	5.87	3.71	3.58
	(en)	5.92	6.02	7.32	7.54	4.02	4.07	5.86	6.13	5.59	5.78	4.60	4.60
Replace	(ch)	4.98	5.47	6.90	6.98	4.15	4.16	5.82	6.23	5.88	6.00	3.71	3.93
	(en)	5.92	5.97	7.32	7.12	4.02	4.26	5.86	6.25	5.59	5.88	4.60	4.54

Table 5: Model performance on AlignBench (Liu et al., 2023a) in the setting of answer replacement and integration.

Model	Acc.	Recall	Precision	F1.
bert-base-chinese (ch)	86.64	86.64	86.68	86.66
bert-uncased (en)	94.98	94.98	94.88	94.91
Multilingual Bert (ch)	86.47	86.47	86.58	86.51
Multilingual Bert (en)	94.73	94.73	94.64	94.67

Table 6: The impact of model selection on detector training.

create a bilingual dataset, we translated the Chinese questions into English. For each response evaluation, we employed an LLM-as-a-judge approach, utilizing the prompt template from AlignBench. The LLM judge then assigned an overall score ranging from 1 to 10 to each LLM response. As indicated in Table 5, both replacement and integration methods significantly enhanced the LLMs’ performance across most datasets. Direct replacement led to more substantial improvements but also introduced a higher rate of errors, as evidenced by the performance dips in GPT-4 and Llama3-Ch. Interestingly, the integration method showed a more pronounced performance improvement in English responses, suggesting that LLMs may possess stronger capabilities for answer optimization in English than in Chinese (Yang et al., 2024a).

The Impact of Different Detection Models. As we build a different low-resource detector for each language, the selection of the tokenizer and classification model would impact the training of the detector thereby influencing the overall performance. We adopt language-specific Bert and multi-lingual Bert models to train our low-resource query detector and report the result in Table 6. As shown in the model, using the language-specific model and tokenizer would slightly improve the result of using a multi-lingual model.

5 Discussion on Other Approach

As the confidence of the generated content is related to its entropy during the generation process, a

natural idea is to calculate the entropy in different languages and compare them to decide which is the best language to answer the question. This approach is widely used for measuring LLMs’ uncertainty and detecting hallucinations (Manakul et al., 2023). However, our trial demonstrates that this approach is infeasible and achieves merely random-guess-level performance when selecting the right language for the given queries.

To explore how to leverage entropy-related statistics to select the target language, we train a model f that takes the statistics as input and outputs the selection of the language Y . The statistics we use for a language l include the entropy of the query E_{Q_l} , the entropy of the response E_{R_l} , the perplexity of the query P_{Q_l} , and the perplexity of the response P_{R_l} . We adopt an MLP as the classification model $f : (E_{Q_l}, E_{R_l}, P_{Q_l}, P_{R_l}) \rightarrow Y$ and train the model on the low-resource query dataset we construct. We trained based on SVM and random forests in Llama2-7b’s output. The accuracy is no more than 60%. This is a merely random-guess-level performance when taking the entropy-related statistics as input. We attribute this to the hallucination issue of LLMs, that the model may become overconfident even with the wrong answer (Groot and Valdenegro-Toro, 2024), which indicates LLMs are not calibrated well now (Zhang et al., 2024).

6 Conclusion

This paper presents a method to improve the multilingual capabilities of LLMs by leveraging knowledge from various languages, which includes a low-resource knowledge detector, a process for selecting languages, and answer replacement & integration. Our experiments show significant enhancements in performance, especially in reducing disparities across languages. Moreover, each module in our method contributes to the improvement. Overall, this study underscores the potential of LLMs to unify multilingual functions and provide insights for future research.

559 Limitations

560 Our method requires training a separate low-
561 resource query detector for each language. This is
562 not convenient as the developer of a certain lan-
563 guage should construct a low-resource training
564 set himself, which involves collecting language-
565 specific data. Also, the dataset should be updated
566 with time with the rise of the new language-specific
567 data.

568 Ethics Statement

569 This study adheres to ethical standards in AI re-
570 search and development. We acknowledge that
571 while our methods aim to enhance the multilingual
572 capabilities of LLMs, they must be implemented
573 with careful consideration of potential biases. Ef-
574 forts were made to ensure that the knowledge ag-
575 gregation does not disproportionately favor any par-
576 ticular language or cultural perspective. We also
577 emphasize transparency in our methodologies and
578 findings to enable scrutiny and replication by the
579 research community. The research was conducted
580 without utilizing any personally identifiable infor-
581 mation, thereby safeguarding privacy and uphold-
582 ing data protection standards. We commit to ongo-
583 ing evaluation of our methods in diverse linguistic
584 settings to address and mitigate any emergent bi-
585 ases or disparities. This research seeks not only to
586 advance technology but also to promote inclusiv-
587 ity and fairness in AI applications across different
588 linguistic and cultural groups. In this paper, we
589 utilized AI tools to aid in writing and coding, ensur-
590 ing that they did not directly contribute to the
591 writing process and that their use adheres to aca-
592 demic standards. Additionally, we ensured that all
593 datasets and benchmarks used in the study comply
594 with their intended purposes and standards.

595 References

596 2024. Llama3-chinese. [https://github.com/](https://github.com/LlamaFamily/Llama-Chinese)
597 [LlamaFamily/Llama-Chinese](https://github.com/LlamaFamily/Llama-Chinese).

598 Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi.
599 2024. Knowledge graphs as context sources for llm-
600 based explanations of learning recommendations.

601 01. AI, :, Alex Young, Bei Chen, Chao Li, Chen-
602 gen Huang, Ge Zhang, Guanwei Zhang, Heng Li,
603 Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong
604 Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang,
605 Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang,
606 Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng
607 Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai,

Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. 608
609

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 610
2019. On the cross-lingual transferability of mono- 611
lingual representations. *CoRR*, abs/1910.11856. 612

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, 613
Terra Blevins, Hila Gonen, Machel Reid, Yulia 614
Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 615
2023. Buffet: Benchmarking large language models 616
for few-shot cross-lingual transfer. *arXiv preprint* 617
arXiv:2305.14857. 618

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, 619
Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei 620
Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, 621
Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, 622
Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, 623
Xuanheng Ren, Chuanqi Tan, Sinan Tan, Jianhong 624
Tu, Peng Wang, Shijie Wang, Wei Wang, Sheng- 625
guang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, 626
Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, 627
Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingx- 628
uan Zhang, Yichang Zhang, Zhenru Zhang, Chang 629
Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang 630
Zhu. 2023. Qwen technical report. *arXiv preprint* 631
arXiv:2309.16609. 632

Du Chen, Yi Huang, Xiaopu Li, Yongqiang Li, 633
Yongqiang Liu, Haihui Pan, Leichao Xu, Dacheng 634
Zhang, Zhipeng Zhang, and Kun Han. 2024. Orion- 635
14b: Open-source multilingual large language mod- 636
els. 637

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin 638
Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, 639
Mianqiu Huang, Zhangyue Yin, Kai Chen, and 640
Xipeng Qiu. 2023. Evaluating hallucinations in chi- 641
nese large language models. 642

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, 643
Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: 644
General language model pretraining with autoregres- 645
sive blank infilling. 646

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez 647
de Lacalle, and Mikel Artetxe. 2023. Do multilingual 648
language models think better in english? 649

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, 650
Vidhisha Balachandran, and Yulia Tsvetkov. 2024. 651
Don't hallucinate, abstain: Identifying llm knowl- 652
edge gaps via multi-llm collaboration. 653

Chujie Gao, Qihui Zhang, Dongping Chen, Yue Huang, 654
Siyuan Wu, Zhengyan Fu, Yao Wan, Xiangliang 655
Zhang, and Lichao Sun. 2024. The best of both 656
worlds: Toward an honest and helpful large language 657
model. 658

Tobias Groot and Matias Valdenegro-Toro. 2024. Over- 659
confidence is key: Verbalized uncertainty evaluation 660
in large language and vision-language models. 661

662	Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 38, pages 18126–18134.	716
663		717
664		718
665		719
666		
667		
668	Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges.	720
669		721
670		722
671		723
672	Taicheng Guo, kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In <i>Advances in Neural Information Processing Systems</i> , volume 36, pages 59662–59688. Curran Associates, Inc.	724
673		725
674		726
675		
676		
677		
678		
679	Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2024. Meta-tool benchmark for large language models: Deciding whether to use tools and which to use.	727
680		728
681		729
682		730
683		731
684	Yue Huang and Lichao Sun. 2024. Fakegpt: Fake news generation, explanation and detection of large language models.	732
685		733
686		734
687	Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.	735
688		736
689		737
690		738
691	Viet Dac Lai, Nghia Trung Ngo, Amir Poursan Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.	739
692		740
693		741
694		742
695		743
696	Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024a. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions.	744
697		745
698		746
699		747
700		
701	Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024b. Cmmlu: Measuring massive multitask language understanding in chinese.	748
702		749
703		750
704		751
705	Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models.	752
706		753
707		754
708		755
709	Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering.	756
710		757
711		758
712		759
713	Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023b. A survey on fairness in large language models. <i>arXiv preprint arXiv:2308.10149</i> .	760
714		761
715		762
	Yuan Li, Yue Huang, Yuli Lin, Siyuan Wu, Yao Wan, and Lichao Sun. 2024c. I think, therefore i am: Benchmarking awareness of large language models using awarebench.	763
		764
	Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024d. Quantifying multilingual performance of large language models across languages.	765
		766
	Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.	767
	Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023a. AlignBench: Benchmarking chinese alignment of large language models.	
	Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023b. AgentBench: Evaluating llms as agents.	
	Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023c. Deid-gpt: Zero-shot medical text de-identification by gpt-4. <i>arXiv preprint arXiv:2303.11032</i> .	
	Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.	
	Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. <i>arXiv preprint arXiv:2303.08896</i> .	
	Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.	
	Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.	
	OpenAI. 2023a. Chatgpt. https://openai.com/product/chatgpt .	
	OpenAI. 2023b. Gpt-4. https://openai.com/gpt-4 .	

768	Chen Qian, Xin Cong, Cheng Yang, Weize Chen,	Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li,	823
769	Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong	Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei	824
770	Sun. 2023. Communicative agents for software de-	Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan	825
771	velopment. <i>arXiv preprint arXiv:2307.07924</i> .	Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei	826
		Huang, and Jun Xie. 2023b. Polylm: An open source	827
		polyglot large language model .	828
772	Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen,	Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao,	829
773	Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and	Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong	830
774	Philip S. Yu. 2024. Multilingual large language	Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi,	831
775	model: A survey of resources, taxonomy and fron-	Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang,	832
776	tiers .	Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian,	833
		Yiwen Zhang, He Zhou, Shaowei Hua Liu, Zhe Zhao,	834
777	Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo	Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang	835
778	Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-	Yang, Kyle Richardson, and Zhenzhong Lan. 2020.	836
779	Yeung Shum, and Jian Guo. 2024a. Think-on-graph:	CLUE: A Chinese language understanding evalua-	837
780	Deep and responsible reasoning of large language	tion benchmark . In <i>Proceedings of the 28th Inter-</i>	838
781	model on knowledge graph .	<i>national Conference on Computational Linguistics</i> ,	839
		pages 4762–4772, Barcelona, Spain (Online). Inter-	840
782	Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and	national Committee on Computational Linguistics.	841
783	Xin Luna Dong. 2024b. Head-to-tail: How knowl-		
784	edgeable are large language models (llms)? a.k.a.	Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang,	842
785	will llms replace knowledge graphs?	Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts	843
		for llms: A survey .	844
786	Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qi-	Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye,	845
787	hui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wen-	and Hanwen Gu. 2024b. A survey on multilingual	846
788	han Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu,	large language models: Corpora, alignment, and bias .	847
789	Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen,		
790	Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao,	Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang,	848
791	Chunyuan Li, Eric Xing, Furong Huang, Hao Liu,	Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang,	849
792	Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao,	Dong Yan, et al. 2023. Baichuan 2: Open large-scale	850
793	Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit	language models . <i>arXiv preprint arXiv:2309.10305</i> .	851
794	Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao,		
795	Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang,	Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu,	852
796	Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi	Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024a.	853
797	Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael	Large language models as optimizers .	854
798	Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu		
799	Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang	Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and	855
800	Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi	Xindong Wu. 2024b. Give us the facts: Enhancing	856
801	Zhou, William Wang, Xiang Li, Xiangliang Zhang,	large language models with knowledge graphs for	857
802	Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan	fact-aware language modeling .	858
803	Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue		
804	Zhao. 2024c. Trustllm: Trustworthiness in large lan-	Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran,	859
805	guage models .	Thomas L. Griffiths, Yuan Cao, and Karthik	860
		Narasimhan. 2023. Tree of thoughts: Deliberate	861
806	InternLM Team. 2023. Internlm: A multilingual lan-	problem solving with large language models .	862
807	guage model with progressively enhanced capabili-		
808	ties .	Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng,	863
		Alexander Ratner, Ranjay Krishna, Jiaming Shen,	864
809	SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vip-	and Chao Zhang. 2023. Large language model as	865
810	ula Rawte, Aman Chadha, and Amitava Das. 2024.	attributed training data generator: A tale of diversity	866
811	A comprehensive survey of hallucination mitigation	and bias .	867
812	techniques in large language models. <i>arXiv preprint</i>		
813	<i>arXiv:2401.01313</i> .	Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang,	868
		Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu,	869
814	Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey,	Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b:	870
815	and Jens Lehmann. 2017. Lc-quad: A corpus for	An open bilingual pre-trained model . <i>arXiv preprint</i>	871
816	complex question answering over knowledge graphs .	<i>arXiv:2210.02414</i> .	872
817	In <i>International Semantic Web Conference</i> , pages		
818	210–218. Springer.	Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Ad-	873
		hikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin	874
819	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	Zhou, Xiang Li, et al. 2023a. Biomedgpt: A uni-	875
820	Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and	fied and generalist biomedical generative pre-trained	876
821	Denny Zhou. 2023a. Chain-of-thought prompting	transformer for vision, language, and multimodal	877
822	elicits reasoning in large language models .	tasks . <i>arXiv preprint arXiv:2305.17100</i> .	878

879	Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the confidence of large language models by eliciting fidelity.	928
880		929
881		930
882		931
883	Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangdong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023b. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.	932
884		
885		
886		
887		
888		
889	Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023c. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models.	
890		
891		
892		
893	A Dataset Details	
894	• TriviaQA (Joshi et al., 2017) is a reading comprehension dataset that features more than 650,000 question-answer-evidence triples. It consists of lots of question-answer pairs created by trivia aficionados, along with independently collected evidence documents—averaging six per question—that offer robust distant supervision for responding to the queries.	
895		
896		
897		
898		
899		
900		
901		
902	• HaluEval (Li et al., 2023a) is a benchmark designed to assess how well LLMs hallucinations—unverifiable or incorrect content in their outputs. It includes a collection of generated texts and human-annotated samples that help evaluate the models’ performance in detecting such errors.	
903		
904		
905		
906		
907		
908	• HalluQA (Cheng et al., 2023) is a dataset consisting of 450 carefully crafted adversarial questions that cover various domains, incorporating elements of Chinese historical culture, customs, and social phenomena. It aims to evaluate LLMs on their propensity to produce two types of errors: imitative falsehoods and factual inaccuracies.	
909		
910		
911		
912		
913		
914		
915	• WebQA (Li et al., 2016) is a large-scale, human-annotated real-world QA dataset, developed to address the scarcity of extensive real-world QA datasets for neural QA systems.	
916		
917		
918		
919	B Experiment Results	
920	We show the full experiment results in Table 7 , Table 8 , and Figure 8 .	
921		
922	C Details of Constructed Dataset	
923	For the generated dataset, inspired by previous studies (Huang et al., 2024; Yu et al., 2023), we employed attribute-guided prompting to instruct LLMs to generate relevant questions on specific topics, as illustrated in Table 9 . We chose GPT-4	
924		
925		
926		
927		
	as our generation model because of its exceptional ability to follow instructions. The prompt template is shown in Figure 9 . For the generated items, we manually evaluate the correctness of its label to ensure the data quality.	
	D Collection of Chinese Domain Dataset	
	Our Chinese domain dataset consists of 227 items. This dataset encompasses knowledge and information specific to Chinese content, including aspects of geography, history, culture, and more. We sourced the content from a broad range of Chinese social media platforms and search engines. After initial curation, we conducted filtering to remove contents that cannot be accurately translated into English or may result in discrepancies in meaning upon translation, such as phrases from ancient Chinese.	
	E Answer Evaluation	
	We adopt the LLM-as-a-Judge for answer evaluation in all experiments to reduce the bias that comes from keyword matching. We use GPT-4 for our evaluation due to its exceptional capability and wide application in previous studies (Liu et al., 2023a; Gao et al., 2024). For the five QA datasets, we use the prompt template shown in Figure 10 .	
	F Prompt Template	
	We show the prompt template used in our study in Figure 9 , Figure 10 , Figure 12 , Figure 11 , Figure 14 , and Figure 13 .	
	G Screenshots of Human Evaluation	
	We show the screenshots of human evaluation in Figure 15 and Figure 16 .	

Dataset	Lang.	ChatGLM3		ChatGPT		GPT-4		Yi-34b		Qwen-turbo		Llama3-Ch.	
		Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.
HalluEval	(en)	18.03%	18.03%	57.98%	57.84%	67.13%	66.99%	42.86%	41.75%	29.31%	29.59%	40.67%	40.67%
	(ch)	11.23%	17.34%	32.07%	52.38%	47.99%	65.05%	25.10%	39.81%	19.35%	26.51%	25.35%	37.33%
HalluQA	(en)	20.00%	25.41%	34.27%	30.90%	51.89%	53.51%	38.38%	47.03%	25.97%	37.57%	22.83%	18.48%
	(ch)	22.16%	22.70%	21.91%	25.28%	49.73%	51.89%	45.95%	45.95%	43.65%	39.78%	15.22%	20.65%
Chinese Domain	(en)	9.52%	21.21%	41.85%	42.73%	56.71%	57.58%	33.33%	57.58%	27.19%	48.25%	30.74%	24.24%
	(ch)	32.47%	25.54%	41.85%	42.29%	59.31%	58.44%	63.64%	57.14%	62.28%	62.28%	18.61%	22.51%
triviaQA	(en)	36.32%	36.38%	90.53%	90.31%	94.09%	93.93%	79.33%	78.90%	59.59%	59.47%	77.27%	77.05%
	(ch)	21.33%	32.22%	54.60%	83.33%	82.77%	92.29%	59.43%	76.27%	41.53%	53.55%	43.92%	66.32%
WebQA	(en)	28.51%	38.96%	59.08%	58.98%	67.70%	69.61%	57.07%	69.71%	49.48%	62.11%	50.00%	47.08%
	(ch)	48.69%	42.07%	57.35%	59.29%	72.52%	72.32%	76.93%	74.12%	71.12%	70.70%	37.02%	40.54%

Table 7: Six LLMs’ performance on the setting without a low-resource detector.

Dataset	Lang.	ChatGLM3		ChatGPT		GPT-4		Yi-34b		Qwen-turbo		Llama3-Ch.	
		Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.	Orig.	Impr.
HalluEval	(en)	18.03%	18.31%	57.98%	57.70%	67.13%	66.57%	42.86%	42.72%	29.31%	29.17%	40.67%	40.25%
	(ch)	11.23%	18.03%	32.07%	56.02%	47.99%	66.16%	25.10%	42.58%	19.35%	28.61%	25.34%	39.97%
HalluQA	(en)	20.00%	25.95%	34.27%	32.02%	51.89%	53.51%	38.38%	46.49%	25.97%	38.67%	22.83%	21.74%
	(ch)	22.16%	23.78%	21.91%	23.60%	49.73%	51.35%	45.95%	44.86%	43.65%	41.44%	15.21%	15.76%
Chinese Domain	(en)	9.52%	32.03%	41.85%	41.85%	56.71%	59.31%	33.33%	61.04%	27.19%	58.33%	30.74%	19.48%
	(ch)	32.47%	30.74%	41.85%	41.41%	59.31%	58.44%	63.64%	60.17%	62.28%	58.33%	18.61%	18.61%
triviaQA	(en)	36.32%	35.78%	90.53%	89.09%	94.09%	93.54%	79.33%	78.73%	59.59%	58.80%	77.27%	76.12%
	(ch)	21.33%	35.94%	54.60%	89.15%	82.77%	93.22%	59.43%	78.18%	41.53%	58.41%	43.92%	74.92%
WebQA	(en)	28.51%	44.38%	59.08%	59.90%	67.70%	70.81%	57.07%	73.72%	49.48%	67.70%	50.00%	46.48%
	(ch)	48.69%	46.99%	57.35%	58.88%	72.52%	71.61%	76.93%	74.22%	71.12%	69.25%	37.02%	41.15%

Table 8: Six LLMs’ performance on the setting without language selection.

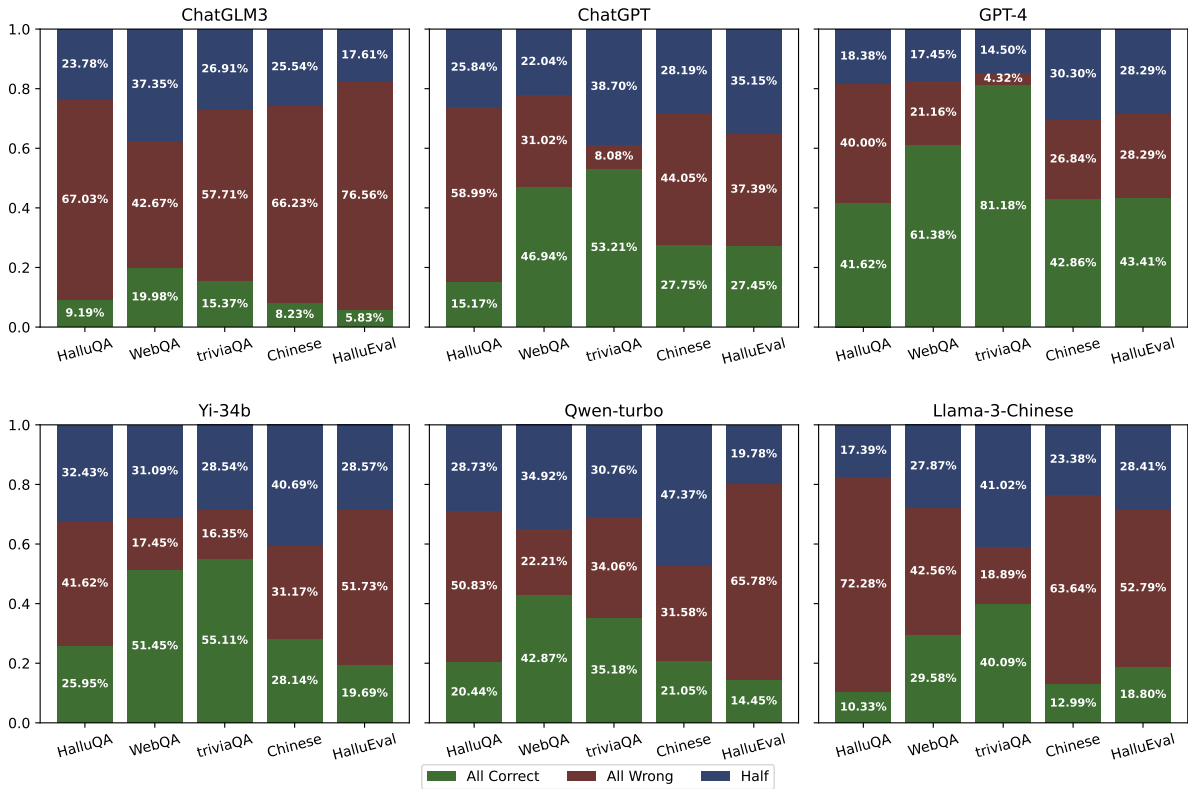


Figure 8: Performance percentage of LLMs across different datasets. ‘All correct’ indicates that the LLMs answered correctly in both the Chinese and English datasets. ‘All wrong’ signifies that the LLMs answered incorrectly in both datasets. ‘Half’ denotes that the LLMs answered correctly in only one of the datasets.

Prompt Template

Next, I will provide you with a topic, and you will assist me in generating data based on this topic. I need you to generate three categories: questions with an English background, questions with a Chinese background, and questions with no specific language.

I will provide you with some examples:

Question: Piaget believes that communication has two functions, one is the egocentric function, and the other is?

Category: English knowledge

Question: With one byte, how many different codes can be generated at most?

Category: Knowledge with no specific language

Question: What are some famous dishes from Guangdong?

Category: Chinese knowledge

For each type of question, you need to generate ten, a total of thirty.

You need to return the data in JSON format, as follows:

```
{
  "Question": "Category",
  "Question": "Category",
  "Question": "Category",
  "Question": "Category",
  ...
}
```

Please generate the corresponding data in Chinese.

The topic I provide is: [TOPIC]

Figure 9: Prompt template for the generated dataset.

Prompt Template

As a helpful assistant, your task is now to help me assess the correctness of the provided answers. I will present a question along with its correct answer. Subsequently, I will also provide you with the answer you need to evaluate. If the answer to be evaluated correctly expresses the same meaning as the correct answer or contains the correct answer, then it is right. Ignore case errors. Although there are some errors in certain explanations within the answer, as long as the core answer is correct, the response is considered correct. Return me only one word: 'correct' or 'wrong'.

Here is the question and its correct answer:

Question: [QUESTION]

Answer: [ANSWER]

Here is the answer you should evaluate: [RES]

Figure 10: Prompt template for LLM-as-a-Judge.

Prompt Template

You are a very helpful assistant. I will provide you with a question and the answers in both Chinese and English. You need to integrate the Chinese and English answers to provide the final answer. During the integration process, you need to follow these rules:

1. You should primarily refer to the Chinese answer, appropriately integrating parts of the English answer.
2. If there is a factual conflict between the English and Chinese answers, you must refer to the Chinese answer.
3. The integrated answer should be of higher quality than the individual answers and better address the corresponding question.
4. The integrated answer must be all in English

Question: [[Q]]

Chinese answer: [[CH_RES]]

English answer: [[EN_RES]]

Figure 11: Prompt template for integration (For the situation when the selected language is English).

Prompt Template

你是一个非常有帮助的助手。我将给你提供一个问题，以及该问题的中英文的答案。你需要融合中英文答案，给出最终的答案。在融合答案的过程中，你需要遵循下面的规则：

1. 你需要着重参考英文的答案，适当融合部分中文的答案。
2. 如果英文的答案与中文的答案发生事实性冲突，你必须参考英文的答案。
3. 融合后的答案应该比融合前的答案具有更高的质量，更好地回答对应的问题。
4. 融合后的答案必须全都是中文。

问题: [[Q]]

中文答案: [[CH_RES]]

英文答案: [[EN_RES]]

Figure 12: Prompt template for integration (For the situation when the selected language is Chinese).

Prompt Template

As a helpful assistant, you need to categorize an English question, considering that the background of this question is not common in an English environment. Therefore, you need to choose the most suitable language for this question. You need to analyze the required language context for the question first, and then tell me at the end which language you think is most suitable to answer the question. The question is as follows:

Figure 13: Prompt template for language selection (For the query in English).

Prompt Template

作为乐于助人的助理，您需要将一个中文问题进行分类，考虑到该问题背景在中文环境中并不常见，因此您需要返回最适合该问题的语言。你需要首先对问题所需要的语言环境进行分析，然后在最后告诉我你返回的最适合回答该问题的语言。问题如下：

Figure 14: Prompt template for language selection (For the query in Chinese).

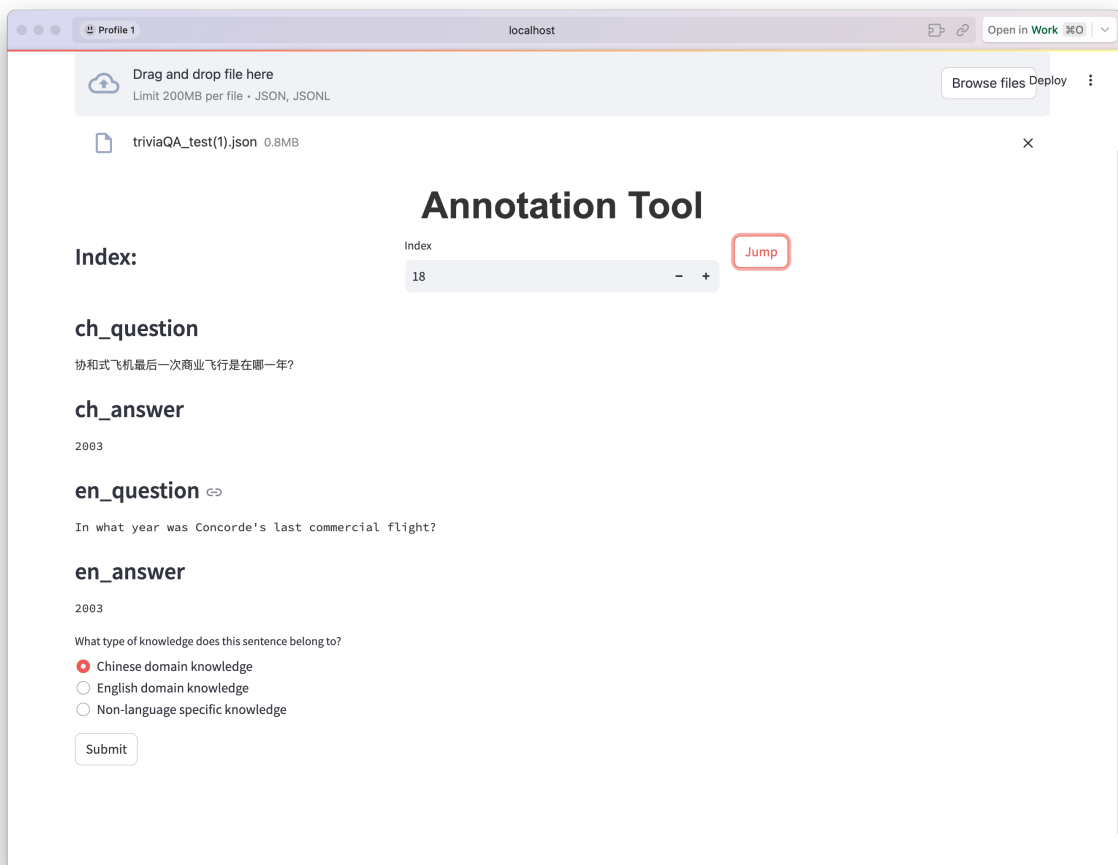


Figure 15: Screenshot of human annotation (1).

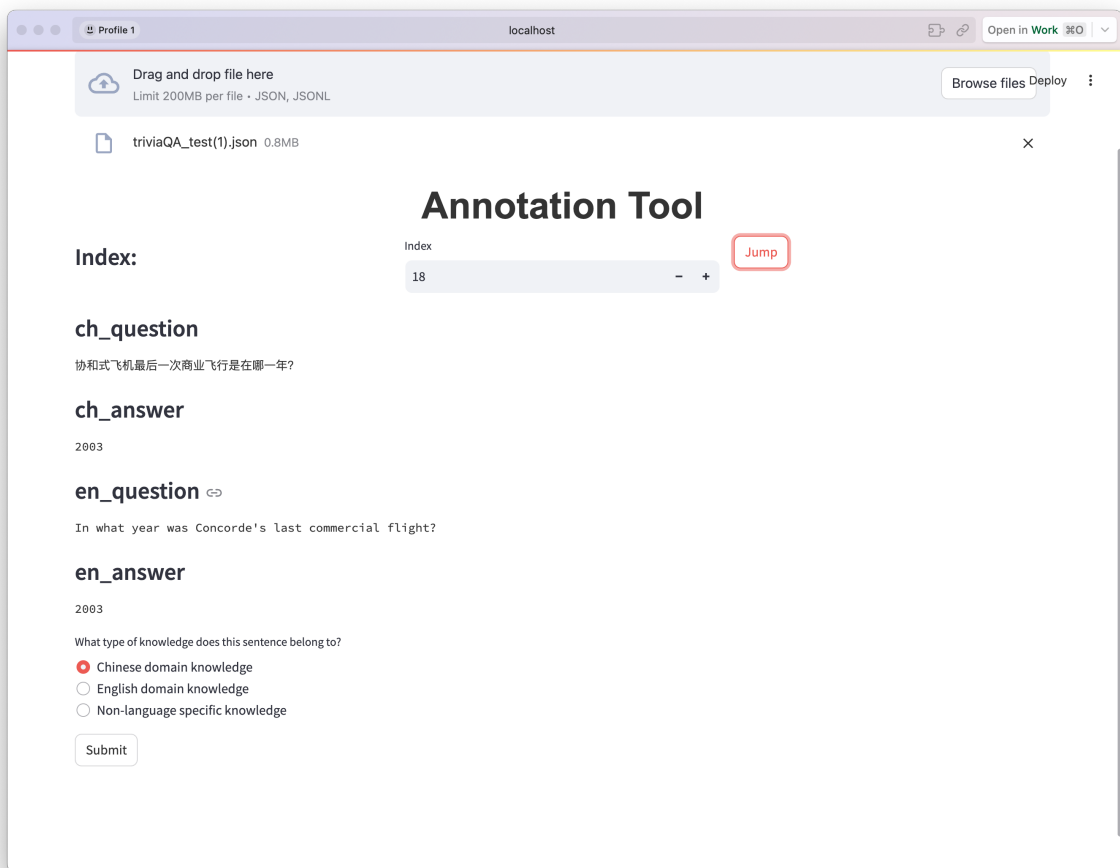


Figure 16: Screenshot of human annotation (2).

Table 9: Topics used for data generation.

<i>Topic Word</i>			
History	Literature	Science	Art
Social Sciences	Technology	Philosophy	Geography
Culture	Health	Artificial Intelligence	Machine Learning
Big Data	Blockchain	Internet of Things	Environmental Protection
Sustainable Development	Energy	Finance	Education
Human Genetics	Artificial Life	Space Exploration	Food Science
Sports	Psychology	Political Science	Economics
Sociology	Law		