Evaluating LLMs' Factual Knowledge Utilization on Unanswerable Questions

Anonymous ACL submission

Abstract

Handling unanswerable questions (UAQ) is crucial for LLMs, as it helps prevent misleading responses in complex situations. While previous studies have built several datasets to assess LLMs' performance on UAQ, these datasets lack factual knowledge support, which limits the evaluation of LLMs' ability to utilize their factual knowledge when handling UAQ. To address the limitation, we introduce a new unanswerable question dataset **FUAQ**, a bilingual dataset with auxiliary factual knowledge created from a Knowledge Graph. Based on FUAQ, we further define two new tasks to measure LLMs' ability to utilize internal and external factual knowledge, respectively. Our experimental results across multiple LLM series show that FUAQ presents significant challenges, as LLMs do not consistently perform well even when they have factual knowledge stored. Additionally, we find that incorporating external knowledge may enhance performance, but LLMs still cannot make full use of the knowledge which may result in incorrect responses.

1 Introduction

Large Language Models (LLMs) have shown strong performance on a wide range of tasks, including logical reasoning and question-answering (Achiam et al., 2023; Wei et al., 2022; Bai et al., 2023). While LLMs demonstrate remarkable performance on traditional question-answering datasets, in real-world applications, questions posed by users may not have definite or factual answers, for example: *"Who is the sibling of Nero Caesar and also the father of Seti I?"*. Specifically, these questions lack factual answers since there is no supporting factual knowledge either in the real world or within the constraints of the user's context. Hence, we refer to them as **unanswerable questions** (**UAQ**)[1] in this paper. When faced with UAQ, if LLMs provide counterfactual responses, they might mislead users and cause unexpected consequences.

Several researchers have built unanswerable datasets (Yin et al., 2023; Hu et al., 2023; Liu et al., 2023) for LLMs evaluation. With the help of these datasets, we can effectively assess the LLMs' ability to discriminate between unanswerable questions (UAQ) and answerable questions (ABQ). However, these datasets have non-negligible shortcomings: **(1) UAQ without explicit factual knowledge support**: The UAQs in the previous studies are mainly sourced from web-crawling (Yu et al., 2022; Yin et al., 2023), brainstorming (Hu et al., 2023), or obtained by replacing key entities in correct sentences with fake ones (Liu et al., 2023). These existing datasets only provide answers or labels without the information of supporting factual knowledge. This makes it difficult to evaluate LLMs' ability to utilize internal or external factual knowledge for handling UAQs. **(2) English only**: To our best knowledge, the existing datasets only support evaluation in English. It might be interesting to know whether the ability can be generalized to other languages.

To overcome the above shortcomings, we introduce a new dataset FUAQ, a bilingual unanswerable question dataset in which each question is accompanied by related factual knowledge. The factual knowledge is from Knowledge Graph in two languages, English and Chinese. We first sample factual triples from Wikidata (Pellissier Tanon et al., 2016), a widely used KG, as factual knowledge. Then we further design question templates for different question types. Based on the factual triples and the templates, we generate UAQs and ABQs. As we have all the detailed information during generation, we can design reasoning clues

[1] We focus on the factual questions in this paper

Dataset	CREPE (Yu et al., 2022)	SelfAware (Yin et al., 2023)	FalseQA (Hu et al., 2023)	UnknownBench (Liu et al., 2023)	**FUAQ (Ours)**
Source	Web	Web	Brainstorming	Rewrite	KG
#Questions	8,466	3,369	4,730	6,323	**13,970**
#UAQs	2,202	1,032	2,365	4,251	**6,985**
#Tasks	1	1	1	1	**3**
Language	EN	EN	EN	EN	**EN & ZH**
Answer or Label	✓	✓	✓*	✓	✓
Factual Knowledge	✗	✗	✗	✗	✓

Table 1: Statistics of unanswerable question datasets (data sampled from other datasets is excluded). **FUAQ** is a large unanswerable dataset with auxiliary **factual knowledge**. Besides, it is the only dataset that supports 3 evaluation tasks in 2 languages. *: provide a feasible response as the answer.

as external knowledge for each question. Finally, we attach related factual triples to each question as auxiliary factual knowledge, which can be used for evaluating LLMs' ability to utilize factual knowledge in handling UAQ.

Based on FUAQ, different strategies can be employed to evaluate LLMs' ability to utilize factual knowledge when handling UAQ. In this paper, we propose three evaluation tasks: one basic task similar to traditional classification tasks in existing datasets, and two new tasks specifically designed for evaluation. (1) *Discriminating between UAQ and ABQ*: a basic task that provides UAQ and ABQ directly to LLMs, evaluating their ability to discriminate them. (2) *Evaluating LLMs' ability to utilize internal factual knowledge in handling UAQ*: if related factual knowledge is stored in LLMs, can they utilize the knowledge efficiently? (3) *Evaluating LLMs' ability to utilize external factual knowledge in handling UAQ*: if related factual knowledge is provided for LLMs in CoT, can they utilize the clues to answer UAQ correctly?

Finally, FUAQ contains 6,985 UAQs and an equal number of ABQs, totaling 13,970 questions. Additionally, we construct 8,686 questions for UAQs' relevant knowledge and 13,970 reasoning clues as external knowledge to support the two tasks for in-depth evaluation. All data are presented bilingually in both English and Chinese. The statistics of relevant datasets are detailed in Table 1. From the table, we can find that FUAQ is the largest unanswerable dataset among them. Besides, with auxiliary factual knowledge and reasoning clues, FUAQ is able to support the in-depth evaluation of using factual knowledge for handling UAQ.

In summary, our contributions are as follows:

• We create a new dataset, FUAQ, to evaluate LLMs' ability to handle the unanswerable questions. The questions are accompanied by factual knowledge from a KG. To our best knowledge, FUAQ is the largest unanswerable dataset among the existing datasets and the first that has auxiliary factual knowledge which makes in-depth evaluation of LLMs possible. Moreover, FUAQ is in two languages, English and Chinese.

• We define two new tasks to comprehensively assess LLMs' ability to utilize internal and external factual knowledge in handling UAQ, respectively. During the evaluation, we design a new metric, knowledge-aware refusal rate, to measure the performance.

• Based on our dataset, we evaluate across multiple series of LLMs. Insights obtained from the evaluation are summarized as follows:
(1) FUAQ is a challenging benchmark for LLMs in discriminating between UAQ and ABQ.
(2) Despite LLMs having stored extensive factual knowledge within their parameters, they fail to effectively utilize internal knowledge in this task.
(3) External factual knowledge may help LLMs to discriminate UAQ and ABQ. However, LLMs still can not make full use of them in handling UAQ.

2 Related Work

2.1 Unanswerable Question Datasets

Existing unanswerable question datasets are built from multiple sources, including web-crowded (Yu et al., 2022; Yin et al., 2023) and brainstorming (Yin et al., 2023). The number of UAQs in these datasets is limited because UAQs are naturally rare in the real world. To scale up the dataset, Liu et al. (2023) build a large synthetic unanswerable question dataset. It collects a list of false entities and constructs UAQs by filling false entities in predefined templates or replacing key entities in ABQs. These datasets focus on evaluating LLMs' ability

to discriminate UAQ and ABQ in English, only providing answers or labels without information of supporting factual knowledge. This makes it difficult to evaluate LLMs' ability to apply internal or external factual knowledge for handling UAQs. Our dataset FUAQ is created from scratch based on a multilingual Knowledge Graph, collecting triples as auxiliary factual knowledge to construct UAQs and ABQs, which makes in-depth evaluations of LLMs possible.

2.2 Evaluation of LLMs' Internal Knowledge

Researchers have proposed several benchmarks to evaluate LLMs' internal knowledge by open-ended generation (Joshi et al., 2017; Paperno et al., 2016; Lin et al., 2021). While the open-ended generation setting assesses LLMs' ability to "speak out" their internal knowledge, it isn't easy to evaluate (Chang et al., 2024). Alternatively, multiple-choice is adopted in many benchmarks as a feasible setting, including MMLU (Hendrycks et al., 2021), C-Eval (Huang et al., 2024), and LogiQA(Liu et al., 2020). Therefore, for the convenience of evaluation, following the setup of previous research, we design the questions for querying knowledge relevant to UAQ in the multiple-choice format.

3 FUAQ

In this section, we introduce our dataset **FUAQ**, a bilingual unanswerable question dataset in which each question is accompanied by related factual knowledge. First, we introduce the question generation procedure (§3.1) for unanswerable questions (UAQ) and answerable questions (ABQ). Then we illustrate three tasks defined on FUAQ (§3.2). Finally, we report statistics information of FUAQ (§3.3).

3.1 Question Generation

As shown in Figure 1, the question generation has 3 steps: *Question Type Definition*, *Factual Triple Sampling*, and *Template Generation & Filling*.

Question Type Definition Inspired by widely used QA datasets (Yih et al., 2016; Gu et al., 2021) and relevant unanswerable question datasets (Hu et al., 2023), we define three question types (QTypes): *Inter*, *Time*, and *Dilemma*. **(1) *Inter***: LLMs need to return the intersection of two non-empty sets, which correspond to the answer sets of two sub-questions. For UAQ, this intersection is an empty set. **(2) *Time***: LLMs need to respond based on the time constraints given in the question. However, such constraints cannot be satisfied for UAQ. **(3) *Dilemma***: LLMs need to answer questions that provide candidate answers, but for UAQ, all provided candidates are incorrect. Table 2 shows the examples for each QType.

Factual Triple Sampling Once the question type is determined, we need to acquire bilingual factual knowledge to construct the questions. We sample factual triples as knowledge from Wikidata (Pellissier Tanon et al., 2016), a reliable and extensive KG that serves as a central repository for structured multilingual factual data across various subjects.

First, we send a query to Wikidata via API [2] to fetch properties[3] and their corresponding descriptions. The property description provides the meaning and usage of the property. Then we define the following criteria to choose properties: (1) it can be easily understood with the help of its description, (2) it appears at least 5 times in Wikidata, and (3) it is capable of providing factual knowledge. Through the aforementioned criteria, we obtained a property list $\boldsymbol{P}_l$, with 724 properties, e.g. *editor* and *cast member*. Finally, we construct different queries corresponding to each QType to retrieve relevant entities with both English and Chinese labels. These entities combine with properties, yielding factual triples that serve as knowledge. Examples and details of factual triple sampling for three QTypes are listed in Appendix A.

Template Generation and Filling Till now, we have factual triples and planned answers to use in the subsequent steps. The next step is to generate bilingual templates and fill relevant information into templates to generate questions. In our approach, we first mask the entities in the relevant factual triples and keep properties unchanged, using *[E_i]* to mask the entities intended to appear in the question and *[Ans]* to mask the position intended as the question target, e.g. *([E_1], editor, [Ans]) & ([E_2], cast member, [Ans])*. Then we provide them to GPT-3.5 along with the property description to generate templates in the target language, e.g. "*editor is the person who checks and corrects a work (such as a book, ... etc.)*". The prompt we use is shown in Appendix E. The question templates generated by GPT-3.5[4] contain errors in some cases,

[2] https://query.wikidata.org/

[3] Property in Wikidata can be interpreted as a relation or an attribute in triples.

[4] gpt-3.5-turbo-0125

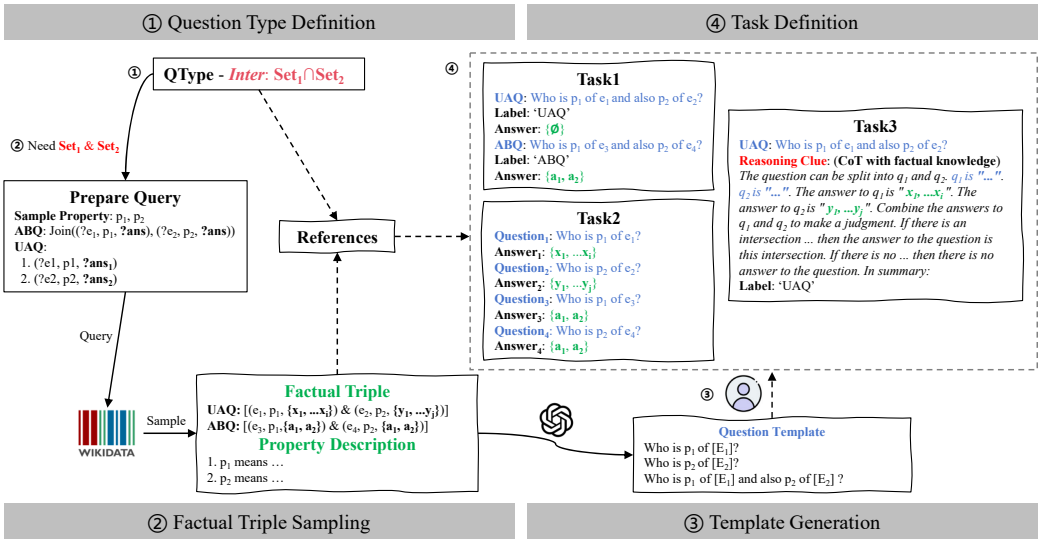

Figure 1: Dataset Construction Process (QType *Inter* in English as an example) for unanswerable question (**UAQ**) and answerable question (**ABQ**): (1) Define the question type. (2) Sample factual triples from Wikidata as knowledge. (3) Generate questions by filling in the templates generated by LLM. (4) Define three tasks and compose unique inputs with information from the preceding steps as references.

QType	Description	Example
Inter	Return intersection of two sets	(UAQ) $\boldsymbol{Q}_{i_1}$: *Who is the **editor of Enneads** and also the **cast member in The Sixth Sense**?*
		(ABQ) $\boldsymbol{Q}_{i_2}$: *Who is the **editor of Die Rote Fahne** and also the **cast member in The Eternal Jew**?*
Time	Consider time constraints	(UAQ) $\boldsymbol{Q}_{t_1}$: *Erfurt was twinned with which city **from 1957 to 1962**?*
		(ABQ) $\boldsymbol{Q}_{t_2}$: *Erfurt was twinned with which city **from 1971 to 1976**?*
Dilemma	Provide candidate answer	(UAQ) $\boldsymbol{Q}_{d_1}$: *What tribe does Segestes belong to, **Mohawk people** or **Khamti people**?*
		(ABQ) $\boldsymbol{Q}_{d_2}$: *What tribe does Segestes belong to, **Khamti people** or **Cherusci**?*

Table 2: Question type (**QType**) of unanswerable question (**UAQ**) and answerable question (**ABQ**) in FUAQ. Examples in Chinese are listed in Appendix B.

including semantic errors, and slots missing. Examples are listed in Table 9. To ensure the quality of the question templates, we manually inspect all templates generated by GPT-3.5, revise or discard incorrect ones, and obtain 864 templates for each language, e.g. "*Who is the editor of [E_1] and also the cast member in [E_2]?*". Finally, we fill in the templates with entities and get the questions. For example, we put "*Enneads*" and "*The Sixth Sense*" into slots *[E_1]* and *[E_2]* of the template and then get the question "*Who is the editor of Enneads and also the cast member in The Sixth Sense?*".

Through the above processes, we sampled factual triples from Wikidata (*["(Enneads, editor, {Porphyry, ...x_i}) & (The Sixth Sense, cast member, {Mischa Barton, ...y_j})]*") and generated corresponding ABQ or UAQ by filling in the template, e.g. "*Who is the editor of Enneads and also the cast member in The Sixth Sense?*" (UAQ Q_{i_1} in Table 2). The answers to these questions are decided as well. For example, "*{Rosa Luxemburg, ...}*" for ABQ Q_{i_2} and "*None*" for UAQ Q_{i_1} in Table 2.

3.2 Task Definition

In this section, we define three tasks for evaluating LLMs' ability to utilize factual knowledge for handling UAQ and introduce their distinctive input drawing from the generated questions and factual triples outlined in §3.1.

Task 1: Discriminating between UAQ and ABQ
In this task, we intend to examine LLMs' ability to discriminate UAQ and ABQ. Following the settings of previous works (Yin et al., 2023; Liu et al., 2023; Amayuelas et al., 2023), we provide questions generated in §3.1 as LLMs' input and attach the corresponding answer (factual answer set to

ABQ and "*None*" to UAQ) for evaluation.

Task 2: Evaluating LLMs' ability to utilize internal factual knowledge in handling UAQ In this task, we first probe LLM's capacity of knowledge relevant to Task 1 questions and then combine it with Task 1 result to evaluate whether LLM can effectively utilize its internal factual knowledge to handle UAQ.

We probe LLMs by asking questions about relevant factual knowledge of UAQ. If LLMs answer correctly, we consider their internal knowledge to be correct; otherwise, we consider them to be incorrect. We construct the input in the form of multiple-choice questions, each offering four options. The following outlines the construction procedure according to QTypes.

• For a UAQ from ***Inter***, we first split the question into two sub-questions and provide four options for each sub-question. For example, Q_{i_1} in Table 2, relevant triples are *(Enneads, editor, ans$_1$)* and *(The Sixth Sense, cast member, ans$_2$)*. We can construct two questions Q_1: "*Who is the editor of Enneads?*" and Q_2: "*Who is the cast member in The Sixth Sense?*". When constructing the options list, we ensure that both *ans*$_1$ and *ans*$_2$ appear in the option lists of the two sub-questions. Subsequently, we sample entities of the same type from Wikidata as incorrect options for Q_1 and Q_2, thus completing their options list with four options.

• For a UAQ from ***Time***, we locate the necessary time boundary for solving it and construct the corresponding multiple-choice question. For example, Q_{t_1} in Table 2, the necessary time boundary is *(Erfurt, twin_city_start_time, 1971)*, demonstrating that the time constraint ("*from 1957 to 1962*") is unfeasible. We can construct the following question: "*When did Erfurt first twin with a city?*". Apart from the gold answer "*1971*" and end time point "*1962*" in the question, we randomly sample two time points, one is earlier than "*1957*" and another is later than "*1962*", e.g. "*1954*" and "*1965*" respectively.

• For a UAQ from ***Dilemma***, after eliminating candidate answers, we utilize the remaining content as the question. For example, Q_{d_1} in Table 2, relevant triple is *(Segestes, tribe, Cherusci)*. "*Mohawk people*" and "*Khamti people*" are provided candidate answers. After elimination, we obtain the question "*What tribe does Segestes belong to?*". When constructing the list of options, we ensure that the gold answer "*Cherusci*" and candidate answers provided in Q_{d_1} are included. Subsequently, we expand the number of options to four following the entity sampling strategy outlined in ***Inter***.

Number of entities	9,021
Number of properties	724
Number of Tasks	3
Number of questions (*EN* | *ZH*)	
Total	13,970
UAQ / *ABQ*	6,985 / 6,985
Inter / *Time* / *Dilemma*	3,428 / 3,882 / 6,660

Table 3: Statistical information of FUAQ. We have English and Chinese versions for each question across three tasks.

Task 3: Evaluating LLMs' ability to utilize external factual knowledge in handling UAQ In this task, we provide LLMs with well-designed reasoning clues as external factual knowledge, evaluating LLMs' ability to utilize external factual knowledge in handling UAQ. A reasoning clue is in the form of CoT: (1) decompose input question Q into several steps (contain question from Task 2), (2) come up with relevant factual knowledge, and (3) answer Q based on the preceding information. In the reasoning clue, decomposed questions and relevant factual knowledge (Steps 1 & 2) are provided.

We take Q_{i_1} in Table 2 as an example, reasoning clue for it is: "***(1)*** *The question can be split into q_1 and q_2. q_1 is "Who ...". q_2 is "Who ...".* ***(2)*** *The answer to q_1 is "Porphyry, ...". The answer to q_2 is "Mischa Barton, ...".* ***(3)*** *Combine the answers to q_1 and q_2 to make a judgment. If there is an intersection ... then the answer to the question is this intersection. If there is no ... then there is no answer to the question.*" More examples are shown in Appendix B.

3.3 Dataset Statistics and Manual Inspection

Detailed information about FUAQ is presented in Table 3. FUAQ has three distinctive features: First, it is a large dataset comprising 13,970 questions. Second, it provides auxiliary factual knowledge for each question, which can support evaluation tasks on factual knowledge application. Third, it supports three tasks across two languages. Beyond one basic task, the other two are new tasks designed to comprehensively assess LLMs' capabilities of using factual knowledge they have. To ensure quality, we conducted manual inspection of FUAQ during and after its construction. The inspection result

Language	English				Chinese			
Metric	Refusal Rate			Acc	Refusal Rate			Acc
Model	$R_{ua}\uparrow$	$R_{ab}\downarrow$	$R_{\Delta}\uparrow$	Acc ↑	$R_{ua}\uparrow$	$R_{ab}\downarrow$	$R_{\Delta}\uparrow$	Acc ↑
Open-sourced LLMs								
Llama3	47.20	**23.38**	23.82	48.99	23.72	**13.77**	9.95	35.89
Mistral0.2	62.15	31.27	30.88	**54.32**	**49.49**	41.57	7.92	19.31
Qwen2.5	**65.74**	34.65	**31.09**	44.29	48.62	31.93	**16.69**	**43.79**
GLM4	63.74	34.73	29.01	49.23	47.03	32.90	14.13	41.53
Average	59.71	31.01	28.70	49.21	42.22	30.04	12.17	35.13
Black-box LLMs								
Gemini-1.5-pro	66.50	**12.25**	54.24	**69.62**	57.42	19.01	**38.40**	**53.98**
GPT-4o-mini	85.05	42.15	42.91	50.51	45.84	**27.70**	18.14	47.02
GPT-4	**85.70**	22.43	**63.26**	66.79	**63.85**	29.33	34.52	51.02
Average	79.08	25.61	53.47	62.31	55.70	25.35	30.35	50.67

Table 4: Refusal rate and Acc of LLMs evaluated in Task 1. R_{ua}, R_{ab}: the refusal rate of UAQ and ABQ respectively. R_{Δ}: R_{ua} - R_{ab}

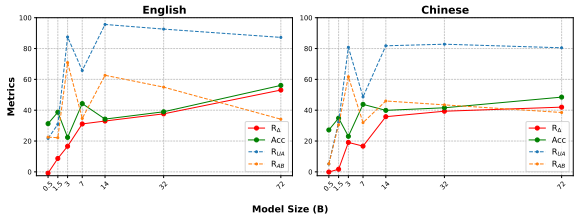

Figure 2: Refusal rate and Acc evaluated in Task 1 of Qwen2.5 series with parameters scaling from 0.5B to 72B. Detailed results are shown in Appendix (Table 10).

shows that 99.2% of questions meet our standards. Details are available in Appendix C.

4 Experiments

4.1 Experiment Setup

We conduct a sequence of experiments by various open-sourced and black-box LLMs, including Llama3 (AI@Meta, 2024), Mistral0.2 (Jiang et al., 2023), Qwen2.5 (Bai et al., 2023), GLM4 (Zeng et al., 2022), Gemini-1.5-pro (Gemini Team, 2024) and OpenAI series (GPT-4o-mini, GPT-4) (Achiam et al., 2023). Open-sourced LLMs we evaluated are corresponding *Chat* or *Instruct* versions around 7B. Our experiments are conducted based on the evaluation framework *lm-evaluation-harness* (Gao et al., 2023), more details can be found in Appendix D.

Metrics

• Following Liu et al. (2023), we obtain ***refusal rate*** using lexical matching by identifying keywords that indicate denial, apology, or abstention. We conducted a human evaluation of sampled LLMs' responses, showing a strong alignment between the lexical matching results and human judgment. Details can be found in Appendix G.1. The refusal rates for UAQ and ABQ are denoted as $\boldsymbol{R}_{ua}$ and $\boldsymbol{R}_{ab}$, respectively. We are also concerned with the difference between these two, denoted as $\boldsymbol{R}_{\Delta}$. The larger the $\boldsymbol{R}_{\Delta}$, the better LLM can discriminate UAQ and ABQ. Ideally, as the capacity of LLM enhances, $\boldsymbol{R}_{ua}$ tends to 1, $\boldsymbol{R}_{ab}$ tends to 0, and $\boldsymbol{R}_{\Delta}$ tends to 1.

• For ABQ in FUAQ, we evaluate the accuracy (***Acc***) of LLM's answer. We search LLM's responses by exact match based on the provided answer list.

• For Task 2, we first report the knowledge pass rate (***KPR***), which measures the percentage of cases in which the LLM successfully passes the knowledge test. To enable a fair comparison of LLMs' ability to utilize internal knowledge across different KPR levels, we introduce a metric called knowledge-aware refusal rate (***KRR***):

$$KRR = (1 + e^{-R_{\Delta} \cdot KPR^{-1}})^{-1} \quad (1)$$

A higher *KRR* indicates better performance, with values ranging from 0 to 1.

4.2 Task 1: Discriminating between UAQ and ABQ

In this task, we examine LLMs' ability to discriminate UAQ and ABQ by directly providing UAQs/ABQs to LLMs, which is close to real-world application scenarios. Additionally, we analyze the relationship between LLMs' parameter size and their performance. The prompts we used are listed in Appendix E.

Results of Task 1 are listed in Table 4. All LLMs have a positive R_{Δ} in two languages. It indicates that LLMs have a certain ability to discriminate UAQ and ABQ when directly confronted with them. However, even the best LLM only achieves 63.26/38.40 R_{Δ} in English/Chinese, which means FUAQ can be a challenging benchmark for evaluating LLMs' ability to discriminate UAQ and ABQ.

In English, black-box LLMs demonstrate superior R_{Δ} compared to open-sourced LLMs. While GPT-4o-mini shows a high refusal rate (85.05) for UAQs, it incorrectly refuses more ABQs than Mistral0.2, leading to lower *Acc* scores (50.51 vs 54.32). Gemini-1.5-pro achieves the highest *Acc* and maintains a R_{Δ} comparable to GPT-4, primarily due to its lower R_{ab}.

We observe similar patterns in Chinese, though with a lower R_{Δ} compared to English. This suggests that Chinese questions pose greater challenges for LLMs in discriminating between UAQ and ABQ. We provide some cases in Appendix

Language	English			Chinese		
Model	**KPR** ↑	**R$_\Delta$** ↑	**KRR** ↑	**KPR** ↑	**R$_\Delta$** ↑	**KRR** ↑
Open-sourced LLMs						
Llama3	65.80	23.82	58.95	52.70	9.95	54.71
Mistral0.2	68.92	30.88	**61.02**	40.21	7.92	54.91
Qwen2.5	**75.16**	**31.09**	60.20	**60.44**	**16.69**	**56.86**
GLM4	65.67	29.01	60.87	55.63	14.13	56.32
Average	68.89	28.70	60.26	52.25	12.17	55.70
Black-box LLMs						
Gemini-1.5-pro	69.03	54.24	**68.69**	76.74	**38.40**	**62.26**
GPT-4o-mini	76.52	42.91	63.66	73.73	18.14	56.12
GPT-4	**81.80**	**63.26**	68.42	**83.21**	34.52	60.23
Average	75.78	53.47	66.93	77.89	30.35	59.53

Table 5: Performance of LLMs in Task 2 knowledge test. **KPR**: knowledge pass rate. **R$_\Delta$**: difference between R$_{ua}$ and R$_{ab}$ (from Table 4). **KRR**: knowledge-aware refusal rate.

G.2.1.

In summary, the above facts indicate that FUAQ is a very challenging benchmark for LLMs. For black-box LLMs, the R_Δ scores range from 42.91 to 63.26 in English and from 18.14 to 38.40 in Chinese, respectively.

Model Scaling

We report *refusal rate* and *Acc* of the Qwen2.5 series, including 7 versions with model scaling from 0.5B to 72B. Figure 2 illustrates the trend of changes in metrics as model scale. Detailed results are listed in Appendix F. As LLMs' parameters scale up, there are noticeable increasing trends in R_Δ. This indicates that **larger LLMs can achieve better results in discriminating UAQ and ABQ**. On the other hand, we observe that R_{ua} and R_{ab} do not show the expected continuous increase or decrease. Instead, they exhibit consistent fluctuating patterns with each other. This suggests that the improvements LLMs achieved in R_Δ do not necessarily lead to better performance in both R_{ua} and R_{ab}: refusing more UAQs while refusing fewer ABQs.

4.3 Task 2: Evaluating LLMs' ability to utilize internal factual knowledge in handling UAQ

In this task, we provide LLMs with questions for knowledge testing and report their knowledge pass rate (*KPR*) and knowledge-aware refusal rate (*KRR*). Results are listed in Table 5.

All black-box LLMs achieve higher KPR than open-sourced LLMs, demonstrating stronger factual knowledge capability. Black-box LLMs also demonstrated a stronger ability in knowledge utilization compared with open-sourced LLMs on average. We observed that while Gemini-1.5-pro fails to achieve the highest KPR, it outperforms all other LLMs in terms of KRR, including GPT-4. Although Gemini 1.5 Pro's KPR in Chinese evaluation still shows a notable gap compared to GPT-4, its superior ability in internal knowledge utilization resulted in its R$_\Delta$ exceeding GPT-4 by 3.88.

Among open-source LLMs, Qwen2.5 leads in KPR for both English and Chinese. Although the English KPR of Mistral0.2 is 6.24 points lower than Qwen2.5, it achieves a higher KRR, suggesting it applies its knowledge more efficiently despite having a smaller knowledge storage. However, Mistral0.2 exhibits a notable decline in KPR when processing Chinese inputs, resulting in a lower R$_\Delta$. These findings highlight that knowledge utilization and knowledge storage are critical determinants of LLMs' overall performance.

When evaluating the same set of Task 2 questions in Chinese versus English, most LLMs show decreased KPR, with Gemini-1.5-pro and GPT-4 being the exceptions, showing increases of 7.71 and 1.41 respectively. However, all models, including those with improved KPR, experience a decline in KRR when processing Chinese inputs, indicating greater difficulty in discriminating UAQ and ABQ in Chinese. Notably, Qwen2.5 and GLM4, despite their lower KPR compared to black-box LLMs, achieve slightly better KRR than GPT-4o-mini. It suggests that Qwen2.5 and GLM4 process a relatively stable knowledge utilization ability across languages. This can be attributed to their strategy during the training phase, where they carefully designed both the data composition and training task to enhance the model's multilingual capabilities.

In summary, the high KPR indicates that LLMs have stored extensive knowledge within their parameters. However, the comparatively low KRR reveals limitations in their ability to effectively utilize internal knowledge. Future research efforts should prioritize developing methods to enhance the utilization of LLMs' internal knowledge, thereby bridging the gap between knowledge storage and practical utilization.

4.4 Task 3: Evaluating LLMs' ability to utilize external factual knowledge in handling UAQ

In this task, we provide questions and CoT with factual knowledge (§3.2) to LLMs (EKnow), aiming to evaluate their ability to utilize external knowledge to correctly address UAQs. Figure 3 shows

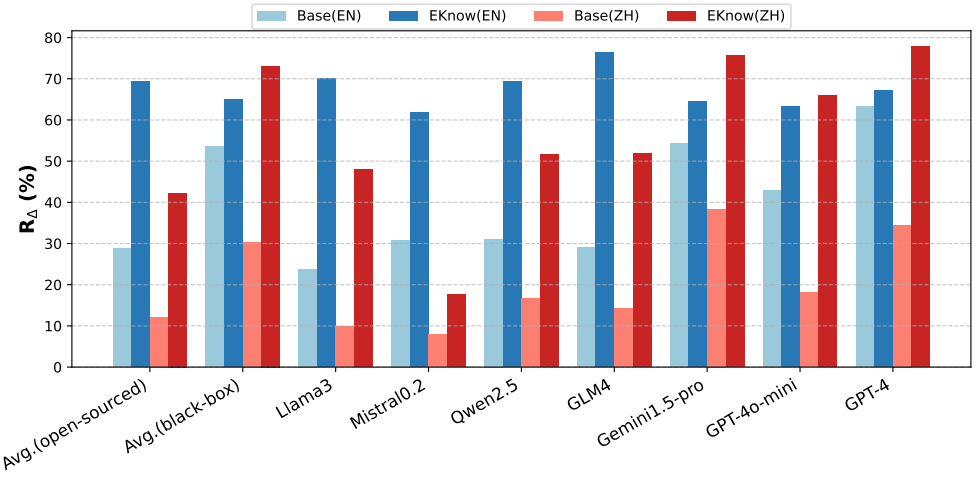

Figure 3: $R_\triangle$ Comparison Between Base and EKnow. *EN* and *ZH* are abbreviations for English and Chinese, respectively. Detailed results are shown in Appendix F.

the $R_\triangle$ of EKnow compared with Task 1 (Base). Detailed results are shown in Appendix F.

With the help of CoT, which provides external reasoning clues with relevant factual knowledge, all LLMs demonstrate improved performance in Task 3 across both languages. Notably, open-sourced LLMs show more substantial gains compared to black-box LLMs, with their relative improvement ($R_\triangle$) exceeding that of black-box LLMs. GLM4 exhibited the most remarkable enhancement, achieving the highest $R_\triangle$ with an improvement of 47.28 (29.01 vs 76.29). These results indicate GLM4's superior capability in leveraging external factual knowledge and effectively handling UAQ in English.

For black-box LLMs, external knowledge does help, but the impact is less significant compared to open-sourced LLMs in English. On one hand, black-box LLMs already achieve good performance by relying solely on their internal knowledge in Task 1, leaving limited room for further improvement. On the other hand, black-box LLMs show an overall decline in both R_{ua} and R_{ab}. We find that black-box LLMs would refuse to respond at a certain rate due to uncertainty about UAQ-related information. With *EKnow* in Task 3, black-box LLMs tend to provide more definitive responses. Some cases are listed in Appendix G.2.2,

In the Chinese setting, the results show different patterns. Black-box LLMs achieve greater improvements, with the average $R_\triangle$ in Chinese even surpassing that in English (73.08 vs 64.96). We also notice that the gap in average $R_\triangle$ between Chinese and English for black-box LLMs in Task 3 is smaller compared to Task 1 (8.12 vs 23.12). This suggests that black-box LLMs demonstrate more balanced capabilities across languages when leveraging external knowledge for UAQ. In contrast, the performance gap widened for open-sourced LLMs (27.13 vs 16.53), including the best-performing GLM4. This indicates that open-source LLMs still have room for improvement in utilizing cross-lingual external knowledge.

In summary, LLMs' ability to effectively utilize external knowledge remains a significant challenge. Even when provided with verified factual knowledge, the best $R_\triangle$ only reaches around 76% in English. The performance tends to decline when using automatic retrieval methods. How to make full use of external knowledge can be an interesting research topic in future work.

5 Conclusion

This paper presents a bilingual dataset FUAQ for deeply evaluating the ability of LLMs to handle unanswerable questions. With auxiliary factual knowledge, FUAQ can support two new tasks other than the classification task. These new tasks can comprehensively assess LLMs' ability to utilize internal and external factual knowledge in handling UAQ, respectively.

Our evaluation results indicate several promising research directions: (1) Enhancing internal knowledge utilization: developing improved methods to activate and utilize factual knowledge already embedded within LLMs; (2) Strengthening external knowledge integration: advancing approaches to better incorporate external knowledge.

Limitations

Following Liu et al. (2023), we use lexical matching to derive the metric *refusal rate*. While human evaluations have confirmed the effectiveness of lexical matching, as evidenced by a Cohen's Kappa of 94.90 (Appendix G.1), there might remain a discrepancy between human evaluation and lexical matching results. Therefore, it is necessary to develop a more precise automated evaluation method in the future.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. 2023. Prudent silence or foolish babble? examining large language models' responses to the unknown. *arXiv preprint arXiv:2311.09731*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From freebase to wikidata: The great migration. In *Proceedings of the 25th international conference on world wide web*, pages 1419–1428.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Xinyan Velocity Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Crepe: Open-domain question answering with false presuppositions. *arXiv preprint arXiv:2211.17257*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

A Details for Factual Triple Sampling

In this section, We provide examples of the factual triple sampling for three QTypes. The templates of SPARQL statements we used to query Wikidata are listed in Table 15.

Denote the property list we obtained in §3.1 as $\boldsymbol{P}_l$, we illustrate the factual triple sampling procedure of three QTypes as follows (take the English version as an example):

Inter We randomly sample two properties (*editor*, *cast member*) from $\boldsymbol{P}_l$ and query with them: *Join((?e_1, editor, ?ans), (?e_2, cast member, ?ans))* (*Select Intersection* in Tabel 15). If the query result exists, e.g. "*[Die Rote Fahne (?e_1), The Eternal Jew (?e_2), {Rosa Luxemburg, ...} (?ans, a non-empty set)]*", we preserve "*[(Die Rote Fahne, editor, ans_1) & (The Eternal Jew, cast member, ans_2]*" as an instance of factual triple and "*{Rosa Luxemburg, ...}*" as the planned answer for ABQ. The ans_1 and ans_2 are full answer set fetched by querying *(Die Rote Fahne, editor, ?ans_1)* and *(The Eternal Jew, cast member, ?ans_2)* respectively (*Select Tail* in Tabel 15). For UAQ, we continue to query with the above properties separately, e.g. query *(?e_1, editor, ?ans_1)* and *(?e_2, cast member, ?ans_2)* by *Select Factual Triples* in Tabel 15. Then we combine triples with no intersection of *?ans_1* and *?ans_2* as factual triples for UAQ, e.g. "*[Enneads (?e1), editor, {Porphyry, ...x_i} (?ans_1, a non-empty set)] & [The Sixth Sense (?e_2), cast member, {Mischa Barton, ...y_j} (?ans_2, a non-empty set, have no intersection with ?ans_1)]*". The planned answer to it is "*None*", signifying that it possesses no answer.

Time We randomly select a property (*spouse*) from $\boldsymbol{P}_l$ and utilize the *Select Time-related Information* query specified in Table 15 to retrieve the relevant factual triples. e.g. *[Queen Paola of Belgium,spouse,?ans] ,[?ans,start time,?time]*. We choose the queried time *?time* as the time constraint for the main question, and the *?ans* as the answer of the ABQ. For UAQ, We randomly sample time points that can not fit the time constraint above, e.g. *[Queen Paola of Belgium, spouse, ?ans]*, *[?and, start time, ?sample time]*. Then by querying the triples with the time, if we obtain an empty result set, we designate this sample time as the time constraint for the UAQ.

Dilemma We choose a property from $\boldsymbol{P}_l$ and use the *select factual triples* query in Table 15 to select factual triples, e.g. *[Russell Banks, personal library at, ?ans]*. Then we query answer set *ans* with *Join(Russell Banks, personal library at, ?ans)* by *Select Tail* in Table 15. We randomly choose an answer from answer set *ans* and use the *Select Options* in Table 15 to select and decide corresponding planned options. All candidates in planned options are incorrect for UAQ, and for ABQ, one candidate in planned options is correct.

B Examples of Data in Three Tasks

Example of Task1 Table 6 shows examples of Task1 questions in Chinese.

Example of Task2 An example of Task 2 evaluation setting is shown as follows:

QType	Description	Example
Inter	Return intersection of two sets	(UAQ) $\boldsymbol{Q}_{i_1}$: 谁既是九章集的编辑，同时也是第六感的演员?
		(ABQ) $\boldsymbol{Q}_{i_2}$: 谁既是红旗报的编辑，同时也是永远的犹太人的演员?
Time	Consider time constraints	(UAQ) $\boldsymbol{Q}_{t_1}$: 1957-1962年期间，爱尔福特的姊妹城市是哪个?
		(ABQ) $\boldsymbol{Q}_{t_2}$: 1971-1976年期间，爱尔福特的姊妹城市是哪个?
Dilemma	Provide candidate answer	(UAQ) $\boldsymbol{Q}_{d_1}$: 桑格斯是哪个部落的成员，莫霍克人还是康迪人?
		(ABQ) $\boldsymbol{Q}_{d_2}$: 桑格斯是哪个部落的成员，谢鲁斯克还是康迪人?

Table 6: Example of Chinese questions.

Task1 UAQ: Who has been one of the head of government of Russian Soviet Federative Socialist Republic and also the father of Ramdas Gandhi?

Task2 Questions (asking UAQ relevant internal knowledge). The gold answer is shown in **bold**:
Q1: Who has been the head of government of Russian Soviet Federative Socialist Republic?
(A) Abdul Kahar of Brunei (B) Mahatma Gandhi **(C) Boris Yeltsin** (D) Salman Khan
Q2: The father of Ramdas Gandhi is?
(A) Boris Yeltsin (B) Syn (C) Lucius Tarquinius Collatinus **(D) Mahatma Gandhi**

KPR=1: Pass knowledge tests, e.g. choosing **(C)** for Q1 and **(D)** for Q2;
KPR=0: Fail at least one knowledge test, e.g. choosing (A)/(B)/(D) for Q1 or choosing (A)/(B)/(C) for Q2;

Example of Task3 Examples of reasoning clues we construct in §3.2 are shown in Table 16 including 3 versions for 3 QTypes. These reasoning clues are used as external knowledge in Task 3.

C Manual Inspection of FUAQ

We perform manual inspection during and after data construction:

1) During data construction, we conduct a comprehensive manual inspection of **all** question templates generated by GPT-3.5. Three annotators independently review each template. Any template marked as incorrect by one annotator is either discarded or revised. Initially, GPT-3.5 generated 1,259 question templates. After our rigorous inspection process, 395 templates are discarded, leaving 864 valid templates. Among these remaining templates, 229 templates are revised. Table 9 presents examples of incorrect question templates originally generated by GPT-3.5 alongside their human-revised versions.

2) After data construction, we conducted a quality assessment by manually examining 900 randomly sampled questions from FUAQ. Our inspection confirms that all question templates are accurately aligned with their designated properties and question types. Additionally, we cross-validated the annotated factual knowledge against Wikipedia and Google to ensure accuracy. In our analysis of 900 questions, we find that only a minimal portion (7 questions, 0.8%) are incorrectly labeled as UAQs due to knowledge error in Wikidata. The vast majority of questions (893, 99.2%) successfully passed our rigorous verification process.

Model Name	**Model Card in HF**
Llama3	meta-llama/Meta-Llama-3-8B-Instruct
Mistral0.2	mistralai/Mistral-7B-Instruct-v0.2
Qwen2.5	Qwen/Qwen2.5-7B-Instruct
GLM4	THUDM/glm-4-9b-chat-hf
Qwen2.5 Series	
0.5B	Qwen/Qwen2.5-0.5B-Instruct
1.5B	Qwen/Qwen2.5-1.5B-Instruct
3B	Qwen/Qwen2.5-3B-Instruct
7B	Qwen/Qwen2.5-7B-Instruct
14B	Qwen/Qwen2.5-14B-Instruct
32B	Qwen/Qwen2.5-32B-Instruct
72B	Qwen/Qwen2.5-72B-Instruct

Table 7: Open-sourced LLMs we evaluate and their corresponding model cards in Hugging Face.

D Experiment Setup

Open-sourced LLMs we evaluate are listed in Table 7. For all 7B LLMs, we set the *temperature* to 0 and inference on V100 with *dtype* set to *float16*. For Qwen series LLMs, we infer those versions smaller than 32B on local following the above setting. For 72B, we fetch the response from API [5].

For Task 1 and Task 3, we set output_type = "*generate_until*" and calculate the refusal rate by a lexical matching function defined by us. For Task 2, we set output_type = "*multiple_choice*" and use "*acc*" as the metric, which is defined by *lm-evaluation-harness*.

[5] https://api.together.ai/

Model	Lexical Matching	Human	Cohen's Kappa
Llama3	35.33	33.22	94.11
Mistral0.2	47.33	41.22	89.42
Qwen2.5	38.89	38.89	96.30
GLM4	57.44	57.33	99.77
Average	44.75	42.67	94.90

Table 8: Refusal Rate obtained by lexical matching and human judgment on sampled data. We apply stratified random sampling to each LLM, drawing a sample of 900 cases based on Qtype and label ("UAQ"/"ABQ"), totaling 3,600 cases. Cohen's Kappa shows a strong alignment between them.

Type	Examples
Semantic Error	✗ What is the function of [E1]'s GPU?
	✓ What type of GPU does [E1] use?
Missing Slot	✗ What type of goods do shops typically sell?
	✓ What type of goods do [E1] typically sell?

Table 9: Incorrect question templates generated by GPT-3.5 and templates after human revision.

E Prompt

Prompt for Template Generation Here is the prompt we used for template generation.

Prompt for Template Generation

Turn the Property into a question template. Take its Description in WikiData as a reference. Return in the following format {"Template":" "}. Some examples are shown below.
...
Property: [E1], editor, [Ans]
Description: editor is the person who checks and corrects a work (such as a book ...)
Template: Who is the editor of [E1]?
...
Property: {*property*}
Description: {*description*}
Template:
==========
将关系转换为问句模板。请以WikiData中对关系的描述作为参考。以{"Template":" "}形式返回。下面是一些示例。
...
关系: [E1], 编辑者, [Ans]
描述: 编辑工作的编辑，如书或定期刊物...
模板: 谁是[E1]的编辑?
...
关系: {*property*}
描述: {*description*}
模板:

Prompt for Evaluation are show in Table 12, which are used in §4 .

Language	English				Chinese			
Metric	Refusal Rate			Acc	Refusal Rate			Acc
Model Size	$R_{ua}\uparrow$	$R_{ab}\downarrow$	$R_{\Delta}\uparrow$	Acc ↑	$R_{ua}\uparrow$	$R_{ab}\downarrow$	$R_{\Delta}\uparrow$	Acc ↑
0.5B	21.72	22.52	-0.80	31.27	5.11	**5.21**	-0.10	27.19
1.5B	30.84	**22.15**	8.69	38.70	32.27	30.55	1.72	34.85
3B	87.49	70.88	16.61	22.36	80.80	61.70	19.10	23.06
7B	65.74	34.65	31.09	44.29	48.62	31.93	16.69	43.79
14B	**95.63**	62.65	32.98	34.24	81.78	45.98	35.79	39.90
32B	92.61	54.97	37.64	38.97	**82.81**	43.46	39.34	41.57
72B	87.23	34.16	**53.07**	**56.09**	80.49	38.51	**41.98**	**48.48**

Table 10: Refusal rate and Acc evaluated in Task 1 of Qwen2.5 series (Detailed results of Figure 2).

Metric	$R_{ua}\uparrow$		$R_{ab}\downarrow$		$R_{\Delta}\uparrow$		Acc ↑	
Model	Base	EKnow	Base	EKnow	Base	EKnow	Base	EKnow
English				**Open-sourced LLMs**				
Llama3	47.20	88.05	23.38	18.00	23.82	70.05	48.99	69.39
Mistral0.2	62.15	86.37	31.27	24.48	30.88	61.89	54.31	76.85
Qwen2.5	65.74	87.19	34.65	17.91	31.09	69.28	44.29	73.61
GLM4	63.74	89.71	34.73	13.42	29.01	76.29	49.24	81.42
Average	59.71	87.83	31.01	18.45	28.70	69.38	49.21	75.32
				Black-box LLMs				
Gemini-1.5-pro	66.50	67.64	12.25	3.22	54.24	64.42	69.62	86.36
GPT-4o-mini	85.05	83.69	42.15	20.34	42.91	63.35	50.51	74.46
GPT-4	85.70	75.06	22.43	7.95	63.26	67.12	66.79	83.71
Average	79.08	75.46	25.61	10.50	53.47	64.96	62.31	81.51
Chinese				**Open-sourced LLMs**				
Llama3	23.72	78.35	13.77	30.37	9.95	47.99	35.89	50.28
Mistral0.2	49.49	91.41	41.57	73.87	7.92	17.54	19.31	6.00
Qwen2.5	48.62	73.16	31.93	21.57	16.69	51.58	43.79	69.76
GLM4	47.03	74.57	32.90	22.71	14.13	51.87	41.53	70.51
Average	42.22	79.37	30.04	37.13	12.17	42.25	35.13	49.14
				Black-box LLMs				
Gemini-1.5-pro	57.42	86.29	19.01	10.67	38.40	75.62	53.98	72.40
GPT-4o-mini	45.84	95.29	27.70	29.42	18.14	65.87	47.02	66.89
GPT-4	63.85	83.58	29.33	5.82	34.52	77.76	51.02	81.36
Average	55.70	88.39	25.35	15.30	30.35	73.08	50.67	73.55

Table 11: Performance of LLMs with the help of external knowledge (Task3). **Base**: Task 1 setting; **EKnow**: external knowledge in Task 3. (Detailed results of Figure 3).

F Detailed Experiment Results

Task1: Parameter Scaling We report detailed results of Task1 (Figure 2) in Table 10.

Task3: Evaluating LLMs' ability to apply external factual knowledge in handling UAQ The *refusal rate* and *Acc* of LLMs with the help of external knowledge (EKnow) are shown in Table 11.

G Case Study and Discussion

G.1 Human Evaluation of Refusal Rate

We apply stratified random sampling to the output of 6 LLMs in Task 1, drawing a sample of 900 cases based on Qtype and label ("UAQ"/"ABQ"), which cumulates to a total of 5,400 cases. After human annotation, we calculate Cohen's Kappa coefficients between the refusal rate obtained by lexical matching and human judgment through the function *sklearn.metrics.cohen_kappa_score*. Table 8 summarizes human evaluation results. Cohen's Kappa coefficients show a strong alignment between the refusal rate obtained by lexical matching and human judgment. This implies that the refusal rate obtained by lexical matching serves as a dependable metric for evaluation.

Setting	Prompt
	English
Task1	Answer the given question in no more than one sentence. Please keep your answer short and concise. Return ##None## if there is no suitable answer. Q: {*question*}. A:
Task 2	The following are multiple choice questions (with answers) Q: {*question*} (A) {*option[0]*} (B) {*option[1]*} (C) {*option[2]*} (D) {*option[3]*} A:
Task 3	Answer the given question in no more than one sentence. Please keep your answer short and concise. Return ##None## if there is no suitable answer. Q: {*question*}. A: {*clue*}
	Chinese
Task1	在1句话以内回答给定问题。确保你的答案简短并简洁。如果你认为不存在合适答案，返回##None##。 问题: {*question*}. 答案:
Task2	下列是多项选择题（有答案） 问题: {*question*}. (A) {*option[0]*} (B) {*option[1]*} (C) {*option[2]*} (D) {*option[3]*} 答案:
Task 3	在1句话以内回答给定问题。确保你的答案简短并简洁。如果你认为不存在合适答案，返回##None##。 问题: {*question*}. 答案: {*clue*}

Table 12: Prompts for evaluation in §4.

G.2 Cases of LLMs' outputs

G.2.1 Cases of Task 1 output

In Table 13, we present the outputs of LLM under Chinese and English inputs. Note that the English input on the left and Chinese input on the right are semantically equivalent. Cases show that it is harder for LLMs to handle inputs in Chinese. Take GPT-4 as an example, it (A) provides a wrong answer that does not fit the time constraint "From 1998 to 2003"; (B) was misled by wrong candidate answers provided by UAQ; (C) fails to consider the constraint, provides a correct answer to question "Who is the mother of Ptolemy XI Alexander II?", which does not fit the other constraint "has been one of the heads of government of Sicily".

G.2.2 Cases of Task 3 output

We present two cases in Table 14. In case (A), Gemini-1.5-pro provides an incorrect answer that is unrelated to the UAQ in Base. Even with external knowledge (EKnow), Gemini-1.5-pro still fails to refuse this UAQ. However, it shows some improvement by providing an answer that is at least relevant to the UAQ (correctly identifying "Wilhelm II" as the spouse of Princess Hermine Reuss of Greiz). In case (B), GPT-4 refuses UAQ in Base but fails to maintain this refusal in EKnow. When we presented GPT-4 with Task 2's question without providing options, it expressed uncertainty about the question. This suggests that GPT-4's refusal in Base stems from its uncertainty about UAQ rather than its judgment based on internal knowledge. In EKnow, GPT-4 provides an incorrect response because EKnow partially addresses some of its initial concerns, leading it to attempt an answer.

Task1 Input	English	Chinese
(A) UAQ	Federico Bonazzoli played for which team from 1998 to 2003?	1998年到2003年期间，费德里科·博纳佐利是哪个团队的成员?
Qwen2.5	✗US Cremona	✗国际足联裁判员 (FIFA Referees)
GPT-4	✓##None##	✗AC米兰 (A.C. Milan)
(B) UAQ	Where was Ahwak recorded, Trident Studios or Olympic Studios?	Ahwak在哪里录制的，三叉戟工作室还是奥林匹克录音室？
Qwen2.5	✗Trident Studios	✗奥林匹克录音室 (Olympic Studios)
GPT-4	✓None	✗奥林匹克录音室 (Olympic Studios)
(C) UAQ	Who has been one of the head of government of Sicily and also the mother of Ptolemy XI Alexander II?	在曾经担任过西西里岛的行政首脑的人中，谁是托勒密十一世的母亲?
Qwen2.5	✓None	✗雷纳托·史济法尼 (Renato Schifani)
GPT-4	✓None	✗克莱奥帕特拉一世 (Cleopatra Selene of Syria)

Table 13: Cases of LLMs' outputs in Task 1.

Model	Example
(A) Gemini-1.5-pro	[Task 1 (UAQ)] Who has been the head of government of Goebbels cabinet and also the spouse of Princess Hermine Reuss of Greiz?
Base	✗Adolf Hitler
EKnow	✗Wilhelm II
(B) GPT-4	[Task 1 (UAQ)] Who has been one of the head of government of Villavaliente and also the mother of Anne Frank?
Base	✓None
EKnow	✗Edith Frank-Holländer
Response of Task 2	This question cannot be answered without additional specific information as Villavaliente does not seem to refer to a known national or municipal government.

Table 14: Cases of LLMs' outputs in Task 3.

Function	**SPARQL**
Select Property	*SELECT ?property ?propertyLabel ?propertyDescription WHERE { ?property a wikibase:Property . OPTIONAL { ?property skos:altLabel ?altLabel . FILTER (lang(?altLabel) = "en") } SERVICE wikibase:label { bd:serviceParam wikibase:language "en" .}}* *SELECT ?property ?propertyLabel ?propertyDescription WHERE { ?property a wikibase:Property . OPTIONAL { ?property skos:altLabel ?altLabel . FILTER (lang(?altLabel) = "zh") } SERVICE wikibase:label { bd:serviceParam wikibase:language "zh" .}}*
Select Property Description	*SELECT ?y WHERE {wd:%prop schema:description ?y. FILTER(LANG(?y) = 'en').}* *SELECT ?y WHERE {wd:%prop schema:description ?y. FILTER(LANG(?y) = 'zh').}*
Select Factual Triples	*SELECT DISTINCT ?x ?y WHERE {?x wdt:%prop ?y .} LIMIT 100*
Select Options	*SELECT ?y WHERE { wd:%qid wdt:P31 ?x. ?y wdt:P31 ?x.} LIMIT 100*
Select Time-related Information	*SELECT DISTINCT ?ans ?start ?end ?point WHERE { wd:%qid p:%prop ?ans. OPTIONAL { ?ans pq:P580 ?start. } OPTIONAL { ?ans pq:P582 ?end. } OPTIONAL { ?ans pq:P585 ?point. } FILTER((BOUND(?start)) || (BOUND(?end)) || (BOUND(?point))). } LIMIT 20*
Select Intersection	*SELECT DISTINCT ?x ?y WHERE {?x1 wdt:%p1 ?y . ?x2 wdt:%p2 ?y .} LIMIT 100*
Select Tail	*SELECT DISTINCT ?y WHERE {wd:%x wdt:%p1 ?y .} LIMIT 100*

Table 15: SPARQL templates for §3.1 Factual Triple Sampling.

Version	Example
English	
Inter	The question can be split into 2 sub-questions, denoted as q1 and q2. q1 is "**The editor of Enneads is?**?". q2 is "**The Sixth Sense's cast members are**?". The answer to q1 is "**Porphyry**". The answer to q2 is "**Mischa Barton, ...**". Combine the answers to q1 and q2 to make a judgment. If there is an intersection of the answers to the sub-questions, then the answer to the question is this intersection. If there is no intersection of the answers to the sub-questions, then there is no answer to the question. In summary, my answer is:
Time	This question can be split into the main question "Erfurt was twinned with which city?" and time "from 1957 to 1962". Through the auxiliary question, "When did Estadio GEBA start participating in association football the first time?", we obtain the START-TIME of the main question sentence "Erfurt was twinned with which city?", the answer of the auxiliary question is "1971". Determined whether "from 1957 to 1962" > START-TIME 1971. If the comparison condition is met, there is an answer, then the corresponding answer will be replied. If the comparison condition is not met, there is no answer to the question. In summary, my answer is:
Dilemma	This question is a dilemma. First we focus on the main problem, then the problem becomes: "What tribe does Segestes belong to, Mohawk people or Khamti people?", and the following options are "Mohawk people or Khamti people". The answer to the previous question is "Cherusci". If the answer appears in the two options, this is the answer, otherwise, there is no answer to the question. In summary, my answer is:
Chinese	
Inter	该问句可被拆分为2个子问题，记为q1和q2。q1是"九章集的编辑是？"。q2是"第六感的演员有？"。q1的答案是"波菲利"。q2的答案是"美莎·芭顿, ..."。结合q1和q2的答案进行判断。如果子问题答案有交集，则该问句的答案为此交集。如果子问题答案没有交集，则该问句没有答案。综上，我的答案是:
Time	该问句是一个关于时间约束的问句，它可被拆分为主问句"爱尔福特的姊妹城市是哪个?"和时间"1957-1962年",通过辅助问句"爱尔福特和杰尔第一次成为友好城市的开始时间？"得到主问句的START_TIME，这个辅助问句的答案是"1908",判断"1900年至1905年"是否与START_TIME "1908"有交集，如果有交集则有答案，则回复对应的答案，如果没有交集，则该问句没有答案。综上，我的答案是:
Dilemma	该问句是一个假两难问题。首先我们忽略候选选项，那么问题变为了："桑格斯是哪个部落的成员，莫霍克人还是康迪人？"，问句中的两个选项是"莫霍克人还是康迪人"。这个问题的答案是"谢鲁斯克"，这些答案是否出现在了后续的选项中？如果出现了那么出现的便是答案，若没有出现则该问句没有答案。综上，我的答案是:

Table 16: Examples of reasoning clues in Task 3.