SKA-Bench: A Fine-Grained Benchmark for Evaluating Structured Knowledge Understanding of LLMs

Zhiqiang Liu[1,3], Enpei Niu[1,3], Yin Hua[1,3], Mengshu Sun[4], Lei Liang[4], Huajun Chen[2,3], Wen Zhang[1,3*]
[1]School of Software Technology, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]ZJU-Ant Group Joint Lab of Knowledge Graph
[4]Ant Group
{zhiqiangliu,zhang.wen}@zju.edu.cn

Abstract

Although large language models (LLMs) have made significant progress in understanding Structured Knowledge (SK) like KG and Table, existing evaluations for SK understanding are non-rigorous (i.e., lacking evaluations of specific capabilities) and focus on a single type of SK. Therefore, we aim to propose a more comprehensive and rigorous structured knowledge understanding benchmark to diagnose the shortcomings of LLMs. In this paper, we introduce ***SKA-Bench***, a **S**tructured **K**nowledge **A**ugmented QA **Bench**mark that encompasses four widely used structured knowledge forms: KG, Table, KG+Text, and Table+Text. We utilize a three-stage pipeline to construct *SKA-Bench* instances, which includes a question, an answer, positive knowledge units, and noisy knowledge units. To evaluate the SK understanding capabilities of LLMs in a fine-grained manner, we expand the instances into four fundamental ability testbeds: *Noise Robustness*, *Order Insensitivity*, *Information Integration*, and *Negative Rejection*. Empirical evaluations on 8 representative LLMs, including the advanced DeepSeek-R1, indicate that existing LLMs still face significant challenges in understanding structured knowledge, and their performance is influenced by factors such as the amount of noise, the order of knowledge units, and hallucination phenomenon. Our dataset and code are available at https://github.com/zjukg/SKA-Bench.

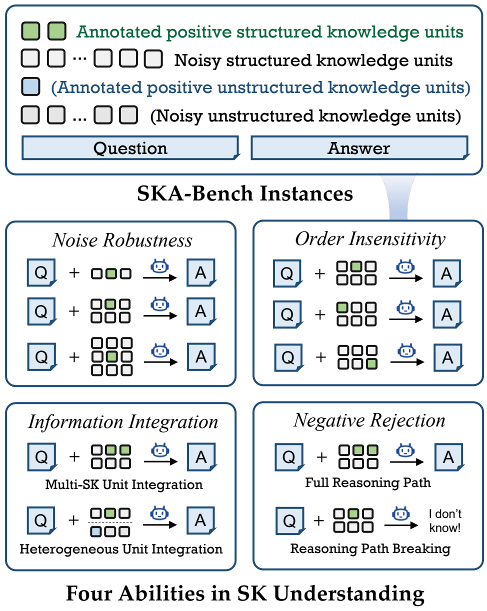

Figure 1: The components of a *SKA-Bench* instance and how to further construct the four ability testbeds for evaluating structured knowledge understanding.

1 Introduction

With the rapid development of large language models (LLMs) (OpenAI, 2023; AI@Meta, 2024), Structured Knowledge (SK), such as knowledge graphs (KG) (Bollacker et al., 2008) and tables, still remain essential due to their systematic and rigorous organizational formats. On the one hand, structured knowledge is usually present in various real-world scenarios (e.g., financial reports with numerous tables (Chen et al., 2021) and product knowledge graphs (Wu et al., 2024)), thus serving as a significant knowledge base for existing LLM systems (Liang et al., 2024; Wang et al., 2025). On the other hand, due to their well-organized structure and intensive knowledge characteristics, structured knowledge is also widely utilized to improve the inference-time performances of LLMs (Li et al., 2024a,b; Guan et al., 2024). Consequently, evaluating the ability of LLMs to understand structured knowledge is a crucial research topic.

Unlike common unstructured text understanding tasks (Guo et al., 2023; Chen, 2024), LLMs still face significant challenges (Fang et al., 2024; Liu et al., 2025) in understanding structured knowledge. This is because LLMs need to capture long-distance contextual dependencies as well as complex relationships and hierarchical structures from

*Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3626–3640
November 4-9, 2025

the given structured knowledge. However, existing benchmarks (Pasupat and Liang, 2015; Wu et al., 2025; Talmor and Berant, 2018; Wu et al., 2024) for evaluating structured knowledge understanding suffer from limitations, including the lack of detailed reasoning path annotations or sufficiently long structured knowledge bases, making it difficult to thoroughly diagnose the shortcomings of LLMs in structured knowledge understanding. Moreover, these datasets primarily focus on single data types, including tables (Pasupat and Liang, 2015; Wu et al., 2025), knowledge graphs (Talmor and Berant, 2018), or hybrid (Chen et al., 2020b; Wu et al., 2024) formats, which restrict their coverage and fail to fully reflect the comprehensive understanding abilities of the models. *Therefore, there is an urgent need for a diverse and fine-grained dataset to comprehensively evaluate LLMs and identify potential bottlenecks in their structured knowledge understanding capabilities.*

To this end, we construct a fine-grained **S**tructured **K**nowledge **A**ugmented QA **Bench**mark, ***SKA-Bench***, which consists of 921 SKA-QA instances and covers four widely used types of structured data. To ensure the quality and complexity of the instances, we propose a novel three-stage construct pipeline for precise positive knowledge unit annotation and the synthesis of long structured knowledge. As illustrated in Fig. 1, *SKA-Bench* instances are composed of a question, an answer, positive knowledge units, and noisy knowledge units, which endow *SKA-Bench* with strong scalability. Ultimately, based on the different compositions of SK units as the given structured knowledge bases, we expand these instances into four distinct testbeds, each targeting a fundamental capability required for understanding SK: *Noise Robustness*, *Order Insensitivity*, *Information Integration*, and *Negative Rejection* for comprehensively diagnosing the shortcomings of LLMs in SK understanding.

We conduct empirical evaluations on 8 representative LLMs. Even advanced LLMs like DeepSeek-R1 continue to face challenges in SK understanding, with their performance significantly influenced by the amount of noise and the order of knowledge units. Moreover, its negative rejection ability is even weaker than that of certain LLMs with 7B parameters. We hope that *SKA-Bench* can serve as a comprehensive and rigorous benchmark to accelerate the progress of LLMs in understanding and reasoning over structured knowledge.

2 Related Work

Evaluation for Structured Knowledge Understanding. Current structured knowledge understanding evaluations often focus on knowledge graphs (Yih et al., 2016; Talmor and Berant, 2018; He et al., 2024) and tables (Pasupat and Liang, 2015; Zhong et al., 2017; Wu et al., 2025). Earlier Table QA datasets, such as WTQ (Pasupat and Liang, 2015), WikiSQL (Zhong et al., 2017), and TabFact (Chen et al., 2020a) require to retrieve several specific table cells with less than 3 hops, posing limited challenges for LLMs. Recently, Wu et al. (2025) propose a more complex Table QA benchmark TableBench for LLM evaluation. However, we believe that the existing evaluations aren't comprehensive enough. On the one hand, the tables in these Table QA datasets are relatively short (average <16.7 rows), making it difficult to evaluate the ability of LLMs to handle long structured knowledge. On the other hand, these datasets lack detailed reasoning path annotations, limiting their utility in fine-grained evaluation of LLMs' understanding capabilities. For existing KGQA datasets, such as WebQSP (Yih et al., 2016), CWQ (Talmor and Berant, 2018), and GraphQA (He et al., 2024), they are constructed upon large-scale KGs, thus providing a foundation for creating long and complex KG understanding datasets. But they also lack precise positive triple annotations for systematic evaluation and analysis.

Evaluation for Semi-structured Knowledge Understanding. To more effectively evaluate the understanding of heterogeneous data, the research community has begun to focus on semi-structured knowledge (Chen et al., 2020b; Zhu et al., 2021; Wu et al., 2024) (i.e., structured data integrated with unstructured textual documents). The semi-structured dataset HybridQA (Chen et al., 2020b), which combines table and textual data, was first proposed. Subsequently, TAT-QA (Zhu et al., 2021) and FinQA (Chen et al., 2021) extend the evaluation of understanding and reasoning to more realistic scenarios based on this data format. In addition, STaRK (Wu et al., 2024) dataset based on KG and textual knowledge bases introduces a new retrieval and reasoning challenge for LLMs. However, these hybridQA datasets are also limited by relatively short length of tables or lack of precise annotations, making them challenging for systematic evaluation.

Based on the above considerations, we believe that offering a diverse, fine-grained, and complex

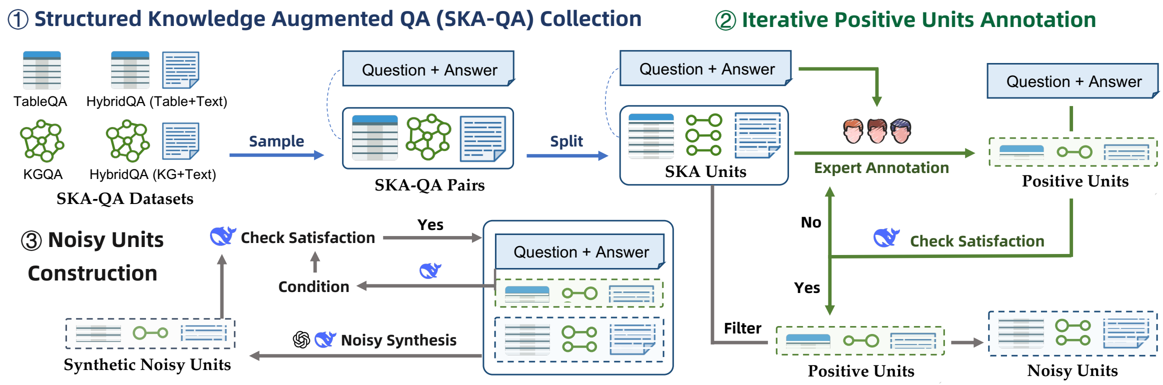

Figure 2: The construct pipeline to generate *SKA-Bench* instance, which consists of structured knowledge augmented question & answer (SKA-QA), positive knowledge units and noisy structured units.

benchmark is valuable for thoroughly evaluating LLMs' structured knowledge understanding ability.

3 SKA-Bench

3.1 Problem Definition

To comprehensively evaluate the ability of LLMs in structured knowledge understanding, *SKA-Bench* incorporates four common types of (semi-)structured data: Knowledge Graph (KG) $\mathcal{G}$, Table $\mathcal{T}$, Knowledge Graph with Textual Documents $\mathcal{G} \cup \mathcal{D}$, and Table with Textual Documents $\mathcal{T} \cup \mathcal{D}$. Following the most existing LLM evaluations (Chang et al., 2024; Guo et al., 2023), *SKA-Bench* also adopts a question-answering (QA) format. For a given question $\mathcal{Q}$ and its corresponding structured knowledge $\mathcal{SK} \in \{\mathcal{G}, \mathcal{T}, \mathcal{G} \cup \mathcal{D}, \mathcal{T} \cup \mathcal{D}\}$, the LLM f_θ aims to generate the correct answer $\mathcal{A}$, such that $\mathcal{A} = f_\theta(\mathcal{Q}, \mathcal{SK})$. We hypothesis that LLMs must accurately understand structured knowledge (SK) as a prerequisite for generating correct answers. Therefore, this task format can thoroughly evaluate the SK understanding capabilities of LLMs.

3.2 SKA-Bench Construction

In this section, we detail the construction process of Structured Knowledge Augmented Benchmark (***SKA-Bench***), which includes three stages: SKA-QA pairs collection, iterative positive units annotation and noisy units synthesis, shown in Fig 2.

3.2.1 SKA-QA Pairs Collections

Knowledge Graph. We randomly select 900 samples from the test set of KGQA datasets: WEBQUESTIONSSP (*WebQSP*) (Yih et al., 2016) and COMPLEXWEBQUESTIONS (*CWQ*) (Talmor and Berant, 2018) as the initial SKA-QA pairs of `KG` subset. These two datasets cover 7 common KG relational patterns (Dutt et al., 2023) and are both based on widely used Freebase KG (Bollacker et al., 2008). For each QA sample, we extract up to 4-hop subgraph of the topic entities (Jiang et al., 2023b) in Freebase as the structured knowledge base.

Table. We randomly select 700 samples from the widely used Table QA dataset *WTQ* (Pasupat and Liang, 2015) and *TableBench* (Wu et al., 2025) with multi-domain, multi-hop question as the initial SKA-QA pairs of `Table` subset. And our selected tables contain at least 6 columns and 8 rows to facilitate the subsequent synthesis of noisy data.

KG with Textual Documents. We choose the *STaRK* (Wu et al., 2024) dataset, which is constructed based on both textual and relational knowledge bases. Specifically, we randomly select 300 QA samples from both *STaRK-Prime* and *STaRK-Amazon*. For each QA sample, we extract the 2-hop subgraph of the answer entity and the textual descriptions of neighboring nodes within subgraph as the corresponding structured knowledge base. Additionally, we remove SKA-QA pairs where the number of triples in subgraph is less than 200.

Table with Textual Documents. For this hybrid data, we also require that QA tasks simultaneously utilize multiple data types. Therefore, we select 200 samples from *HybridQA* (Chen et al., 2020b) dataset as a subset. This dataset necessitates reasoning based on heterogeneous knowledge sources and has been widely used in the research community (Rogers et al., 2023; Fang et al., 2024).

After obtaining the above four types of SKA-QA pairs, we perform a fine-grained split for structured knowledge. Specifically, we regard the triples $\mathcal{F}$ in the KG $\mathcal{G}$ and the rows $\mathcal{R}$ in the tables $\mathcal{T}$ into individual "*structured knowledge units*", represented as $\mathcal{G} = \{\mathcal{F}_i\}_{i=1}^n$ and $\mathcal{T} = \mathcal{H} \cup \{\mathcal{R}_j\}_{j=1}^n$. For

Question: Which nation has the Alta Verapaz Department and is in Central America?
Answer: Guatemala
Positive SK Units:
(Guatemala, location.country.administrative_divisions, Alta Verapaz Department),
(Central America, location.location.contains, Guatemala),
(Guatemala, common.topic.notable_types, Country)
Noisy SK Units:
(Panzós, location.location.containedby, Alta Verapaz Department),
(Central America, location.location.contains, Gran Colombia), ...

Instance in KG subset

Question: Which date the colorado state team scored no points?
Answer: September 20, 1997
Table header:
|Date|Site|Winning team|Winning team score|Losing team|Losing team score|Series|
Positive SK Units:
|September 20, 1997|Fort Collins|Air Force|24|Colorado State|0|AFA 21–14–1|
Noisy SK Units:
|September 6, 1980|Fort Collins|Colorado State|21|Air Force|9|AFA 11–7–1|,
|October 3, 1981|Colorado Springs|Air Force|28|Colorado State|14|AFA 12–7–1|, ...

Instance in Table subset

Question: I have nail dystrophy and chemosis. What skin disease might I have??
Answer: toxic epidermal necrolysis
Positive SK Units:
(Chemosis, phenotype_present, toxic epidermal necrolysis),
(Nail dystrophy, phenotype_present, toxic epidermal necrolysis)
Noisy SK Units:
(EP300, expression_present, cardiac atrium), (PARP1, ppi, DNMT1), ...
Positive unSK Units:
… - mondo_definition: Toxic epidermal necrolysis (TEN) is an acute and severe skin disease with clinical and …
Noisy unSK Units:
- name: breast adenocarcinoma - type: disease - mondo_definition: A carcinoma that arises from glandular epithelial cells of the breast. … …

Instance in KG+Text subset

Question: What are the goals of the athlete who initiated his management career as a player-manager with Middlesbrough in 1994 ?
Answer: 46
Table header: |Name|Years|Apps|Goals|Position|
Positive SK Units: |Bryan Robson|1974-81|249|46|Central midfielder|
Noisy SK Units:
|Billy Bassett|1886-99|311|77|Outside right|,
|Jesse Pennington|1903-22|496|0|Left back|, ...
Positive unSK Units:
Bryan Robson: … Robson began his management career as a player-manager with Middlesbrough in 1994 , retiring from playing in 1997 . …
Noisy unSK Units:
West Bromwich Albion Football Club (/ˈbrɒmɪdʒ, -ɪtʃ/) is an ... …

Instance in Table+Text subset

Figure 3: Four Instances from different subsets of ***SKA-Bench***: LLMs need to understand structured knowledge, then select relevant knowledge units to get the answer.

the table header $\mathcal{H}$, they are separated out independently to preserve the semantic integrity of the table. As for the textual data, we retain the original paragraph-level split in the initial SKA-QA pairs.

3.2.2 Iterative Positive Units Annotation

We invite three human experts with computer science backgrounds to perform positive units annotation. Specifically, we require the human experts to accurately identify the positive units required to derive the answer to the given question. Furthermore, the annotation process need to adhere to the following requirements: **(1)** if the answer is wrong, delete the sample directly; **(2)** if the question involves multiple answers, all positive units require to obtain the answers should be annotated; **(3)** for the `Table` subset and `Table+Text` subset, if the question needs to perform numerical analysis on the entire table, the corresponding SKA-QA pairs should either be removed or the question should be modified; **(4)** if the tables in the `Table` subset and `Table+Text` subset are order-dependent (i.e., modifying the row order would result in semantic errors in the table), this sample should be removed; **(5)** for the `KG+Text` subset and `Table+Text` subset, if question only utilizes one type of knowledge source, the question should be modified or removed.

After each round of annotation, we query the LLM (utilizing DeepSeek-v3 (DeepSeek-AI, 2024)) to determine whether annotated positive units can derive the answer to the given question. If the response is "*No*", re-annotation is performed. The iterative annotation process continues until more than 95% of the samples receive a "*Yes*" response, at which point the iteration is terminated.

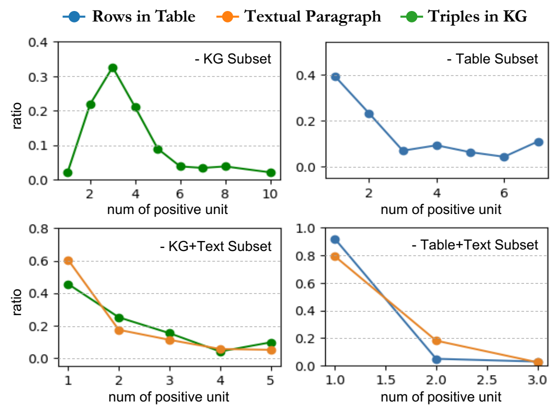

Figure 4: The distribution of the number of positive units across four *SKA-Bench* subsets.

3.2.3 Noisy Units Construction

For `KG` subset and `KG+Text` subset, we regard all knowledge units in the knowledge base except for the positive units as noisy units. The raw tables in the `Table` subset and `Table+Text` subset are typically short (average <17.9 rows), making it hard to comprehensively evaluate the table knowledge understanding of LLMs. Therefore, we introduce an automated noisy data synthesis process as follows.

First, we leverage LLMs with existing SKA-QA instances to generate noisy units. To ensure the diversity of synthesized units, we alternately use GPT-4o (OpenAI, 2023) and DeepSeek-v3 (DeepSeek-AI, 2024) during this process. Meanwhile, we also need to ensure that the synthesized noisy units do not affect the correctness of the answers. To achieve this, we prompt LLM (utilizing

Subset	#avg $\mathcal{Q}$ token	#avg $\mathcal{A}$ num	#num P (SK/unSK)	#avg P token	#num N	#data	Expert Time
SKA-Bench-`KG`	15.75	1.96	4.25	16.77	4541.39	233	5.9 min
SKA-Bench-`Table`	23.31	1.10	3.40	30.88	1521.83	295	3.6 min
SKA-Bench-`KG+Text`	30.76	1.86	2.53/1.92	22.31/1053.55	417.29/79.84	195	6.8 min
SKA-Bench-`Table+Text`	22.41	1.01	1.17/1.28	28.37/203.78	1144.55/661.90	198	5.8 min

Table 1: The data statistics of four subsets in *SKA-Bench*. '#num P' and '#num N' refer to the average number of positive units and noisy units. And '#avg P token' denotes the average number of tokens in positive units. '#data' refers to the numbers of instances in each subsets. The calculation of tokens is based on GPT-4o's tokenizer. 'Expert Time' refers to the median time for each question spent on annotation by human experts.

DeepSeek-v3) with QA and positive units to derive the "*conditions*" that must be satisfied by the rows for answering the question. LLM then verifies whether the generated noisy units meet these "*conditions*". If the response is "*Yes*", the noisy units need to be re-generated by LLMs. After the noise synthesis process, three human experts conduct a manual review of `Table` subset and `Table+Text` subset to evaluate whether the synthetic noise is unsafe and affect the original answers. The review results show that the accuracy rate is 92.5%, and erroneous noise has been deleted.

3.3 Dataset Statistic

Through the aforementioned construction pipeline, we have completed constructing *SKA-Bench* instances as shown in Fig. 3, which consist of four main components: question, answer, positive knowledge units, and noisy knowledge units. Detailed statistics are presented in Table 1. Additionally, we detail the human annotation results, i.e., the number of positive units across the four subsets, as shown in the Fig. 4.

3.4 Testbeds Construction

As shown in Fig. 1, inspired by Chen et al. (2024) in text understanding evaluation, we construct the four testbeds based on *SKA-Bench* instances to evaluate the following fundamental capabilities of LLMs in structured knowledge (SK) understanding:

• **Noise Robustness.** Here, we define noise as the remaining triples in the KG subgraph or the irrelevant rows in the table. We incorporate noise units of varying proportions into the positive knowledge units as the knowledge base to evaluate whether the LLM can robustly provide accurate answers. Considering the differences in the token counts across different knowledge units, we use the total token length as the split standard to construct test sets. Specifically, we construct four test sets {1k, 4k, 12k, 24k} for the `Table` and `KG` subsets, and three test sets {4k, 12k, 24k} for the `Table+Text` and `KG+Text` subsets, with the detailed statistics shown in Table 2. Additionally, to eliminate the influence of the knowledge unit order, we randomly shuffle the SK units in the KG and text units with a random seed of 42, while preserving the original order of the SK units in the Table.

Subset	#num SK	#num unSK	#token
SKA-Bench-`KG-1k`	34.23	-	637.64
SKA-Bench-`KG-4k`	150.34	-	2831.40
SKA-Bench-`KG-12k`	604.35	-	11394.18
SKA-Bench-`KG-24k`	1167.19	-	22036.82
SKA-Bench-`Table-1k`	29.39	-	777.45
SKA-Bench-`Table-4k`	130.39	-	3268.45
SKA-Bench-`Table-12k`	488.00	-	12054.15
SKA-Bench-`Table-24k`	958.78	-	23595.51
SKA-Bench-`KG+Text-4k`	11.37	2.91	3172.54
SKA-Bench-`KG+Text-12k`	40.84	6.79	7417.67
SKA-Bench-`KG+Text-24k`	153.45	19.11	21644.20
SKA-Bench-`Table+Text-4k`	25.82	14.58	3510.74
SKA-Bench-`Table+Text-12k`	75.81	119.01	11899.54
SKA-Bench-`Table+Text-24k`	165.81	369.01	23070.81

Table 2: The data statistics for subsets with different scales of structured knowledge (SK) bases. '#num SK' represents the number of structured knowledge units, '#num unSK' represents the number of unstructured knowledge units in hybrid subsets. And '#token' represents the total number of tokens in the knowledge bases.

• **Order Insensitivity.** SK representation naturally does not depend on any specific order. And in retrieval-augmented scenarios (Fan et al., 2024), the order of retrieved knowledge units tends to be disrupted. Therefore, we expect LLMs to be order-insensitive when understanding SK and capturing the semantic relationships between SK units. In this testbed, we provide SK bases with different permutations of SK units to test whether the LLM is sensitive to order. For SK units in KG and textual units, we position the positive knowledge units at the beginning, randomized positions, and the end of the knowledge base, denoting them as {*prefix, random, suffix*}. For SK units in Table, we additionally introduce the original table order, denoted as {*original*, *prefix*, *random*, *suffix*}. Furthermore, we standardize the test sets to a scale of 4k tokens for `Table` and `KG` subsets, and 12k for `Table+Text`

Model	`KG`				`Table`				`KG+Text`			`Table+Text`		
	1k	4k	12k	24k	1k	4k	12k	24k	4k	12k	24k	4k	12k	24k
Open Source LLMs														
Llama3.1-8B	67.53	58.19	45.86	42.34	27.56	23.52	22.16	13.05	67.02	58.89	49.28	30.27	18.44	12.48
TableGPT-2	78.93	**66.76**	**53.14**	48.49	24.40	24.05	20.09	16.02	64.84	55.16	46.92	35.91	25.63	25.60
Qwen2.5-7B	72.45	60.00	47.98	40.97	**36.69**	**32.04**	**30.45**	**28.68**	**76.51**	62.82	51.83	**38.49**	**36.00**	28.56
GLM4-9B	**82.95**	66.04	52.75	**49.95**	19.55	17.71	16.77	17.26	75.39	65.14	**55.29**	32.13	33.65	**30.13**
Mistral-7B	59.04	60.34	47.98	45.20	17.67	18.11	16.91	16.19	69.37	**66.97**	53.54	29.21	25.40	15.83
Advanced General-Purpose LLMs														
DeepSeek-v3	85.06	73.93	65.85	59.08	54.42	51.83	47.58	45.57	77.12	74.96	68.87	55.64	53.61	48.55
GPT-4o	85.33	73.42	63.04	58.61	51.39	45.18	40.55	38.24	77.38	73.53	67.39	56.52	53.28	51.97
DeepSeek-R1	**89.95**	**81.58**	**70.32**	**64.67**	**61.96**	**61.88**	**61.02**	**58.24**	**83.14**	**78.67**	**71.92**	**62.24**	**57.62**	**56.97**

Table 3: Detailed results of noise robustness analysis. The best results are marked **bold** and the second-best results are underlined in each column. Cells with darker colors indicate the better performance under this subset.

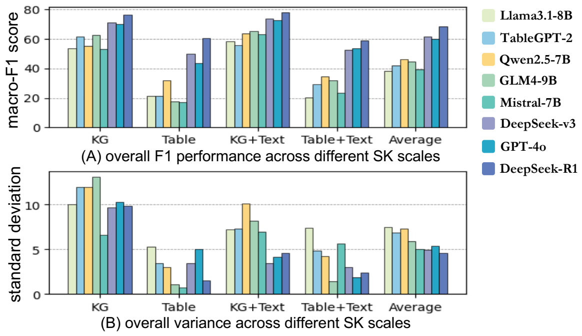

Figure 5: Overall noise robustness results on four subsets. 'Average' represents the average results across on all results of four subsets.

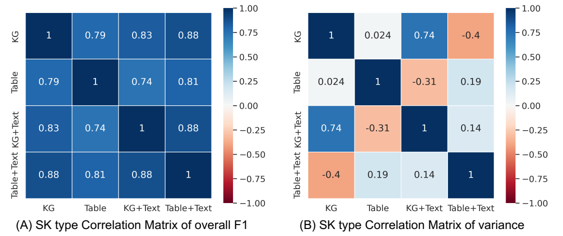

Figure 6: Correlation coefficients of overall F1 and standard deviation across 4 SK types under noise robustness testbed.

and `KG+Text` subsets.

• **Information Integration.** This ability requires LLMs to integrate multiple knowledge units to answer questions, including the integration of multiple SK units and the integration of heterogeneous data (SK+Text) units. Therefore, this testbed focuses on analyzing the performance of LLMs under these two settings. Specifically, we divide our dataset based on the number of knowledge units required to answer each question {2, 3, 4, more than 4} to evaluate the information integration capability of LLMs. Regarding dataset scale and order, we standardize the test set to a scale of 4k tokens for the `Table` and `KG` subsets, and 12k tokens for `Table+Text` and `KG+Text` subsets. Meanwhile, we randomly shuffle (with random seed 42) the SK units in `KG` and text units while preserving the original order of SK units in the `Table` subset.

• **Negative Rejection.** We hope LLMs should minimize the occurrence of hallucination phenomena (Huang et al., 2023) as much as possible when understanding SK. To evaluate this, we construct a negative rejection testbed, where the input SK base consists solely of noisy knowledge units. In this scenario, the LLMs are expected to respond with "*I don't know*" or other rejection signals. In this testbed, the provided SK don't contain any positive units, ensuring broken reasoning paths to evaluate the refusal capability of LLMs. The dataset size and the ordering of knowledge units follow the same settings as "*Information Integration*" testbed.

4 Experiments

4.1 Experimental Settings

Models. Our evaluation is based on popular large language models (LLMs) with a context window of at least 24k tokens. Our evaluated LLMs include advanced general-purpose LLMs: DeepSeek-v3 (DeepSeek-AI, 2024), GPT-4o (OpenAI, 2023), DeepSeek-R1 (DeepSeek-AI, 2025) and common open-source LLMs: Llama-3.1-8B-Instruct (AI@Meta, 2024), Qwen2.5-7B-Instruct (Team, 2024), GLM4-9B-Chat (Zeng et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a). Moreover, we also evaluate the table-specific open-source LLM TableGPT-2 (Su et al., 2024), which are trained based on Qwen2.5-7B.

Evaluation Metric. To evaluate *SKA-Bench*, we utilize the macro-F1 score as our metrics, which measures the agreement between the predicted answer list and the gold answer list. For the negative

Model	KG			Table				KG+Text			Table+Text			
	prefix	*random*	*suffix*	*original*	*prefix*	*random*	*suffix*	*prefix*	*random*	*suffix*	*original*	*prefix*	*random*	*suffix*
Open Source LLMs														
Llama3.1-8B	55.07	58.19	65.85	23.52	22.71	19.47	24.57	61.85	58.89	62.55	18.44	22.37	18.53	24.41
TableGPT-2	**82.07**	**66.76**	77.36	24.05	26.40	17.25	21.62	57.47	55.16	54.53	25.63	36.75	24.44	28.03
Qwen2.5-7B	78.60	60.00	75.70	**32.04**	**33.26**	**24.07**	**31.38**	64.89	62.82	67.74	**36.00**	**48.29**	**29.37**	**37.46**
GLM4-9B	81.30	66.04	**82.55**	17.71	21.15	12.38	16.22	**70.05**	65.14	**69.34**	33.65	41.20	23.89	31.14
Mistral-7B	73.28	60.34	64.30	18.11	21.32	14.76	15.92	63.19	**66.97**	66.36	25.40	33.16	15.44	27.78
Advanced General-Purpose LLMs														
DeepSeek-v3	84.40	73.93	87.52	51.83	49.32	44.75	51.31	76.81	74.96	76.41	53.61	55.80	47.02	49.06
GPT-4o	81.75	73.42	83.69	45.18	45.62	40.47	43.33	74.88	73.53	74.98	53.28	54.88	47.72	52.23
DeepSeek-R1	**89.90**	**81.58**	**89.40**	**61.88**	**67.11**	**61.63**	**64.36**	**79.60**	**78.67**	**81.12**	**57.62**	**59.28**	**53.04**	**57.97**

Table 4: Results of order insensitivity analysis. The best results are marked **bold** and the second-best results are underlined in each column. Cells with darker colors indicate the better performance under this subset.

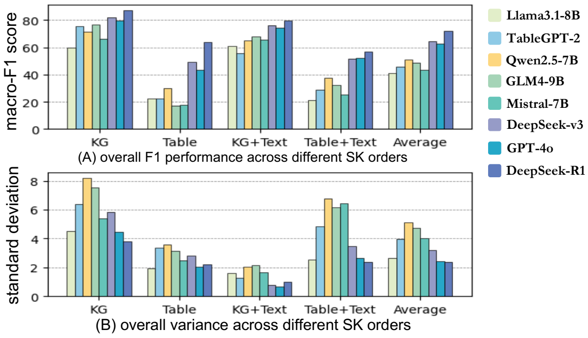

Figure 7: Overall order insensitivity results on four subsets. 'Average' represents the average results across on all results of four subsets.

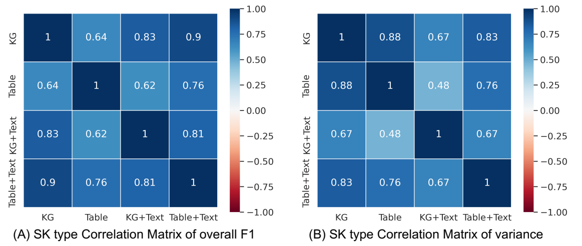

Figure 8: Correlation coefficients of overall F1 and standard deviation across 4 SK types in order insensitivity testbed.

rejection testbed, we adopt the "Rejection Rate" as the evaluation metric, which reflects the proportion of instances where the LLMs provide a refusal response out of the total number of test samples when only noisy knowledge units are provided.

4.2 Noise Robustness Analysis

From the results in Table 3, it can be observed that as the length of SK input to LLM increases, the performance degradation across various LLMs becomes significantly pronounced. In particular, Llama3.1-8B exhibits a dramatic decline of up to 58.77% when evaluated on the `Table+Text` subset from 4k to 24k scale. DeepSeek-R1 demonstrates optimal results across all subsets, whereas GLM4-9B and Qwen2.5-7B achieve relatively competitive performance among the smaller models with the 7-10B parameters.

To further analyze model performance on different data types, we present the mean and standard deviation of F1 scores, and their correlation matrix across 4 subsets, as shown in Fig. 5 and 6. We can observe that the performance trends of different LLMs across 4 SK types are similar in general, with all spearman $\rho > 0.64$. However, there are significant differences in the noise robustness of different LLMs across 4 SK types as shown in Fig. 6(B). GLM4-9B can perform well on the `KG` subset but struggles to understand Table data, and TableGPT-2 leverages large-scale table-related task instruction fine-tuning on the base model Qwen2.5-7B, but its performance on both the `Table` and `Table+Text` subsets is less satisfactory. We attribute this to the loss of generalization capabilities due to its specialized training, making it less adaptable to unseen table formats and other data modalities. Furthermore, we observe that DeepSeek-R1 achieves the lowest average standard deviation, exhibiting the strongest noise robustness. **This suggests that current LLMs are evolving towards greater robustness against noise.**

4.3 Order Insensitivity Analysis

From the results in Table 4, we can observe that when the positive units are concentrated in the prefix or suffix of the structured knowledge base, models tend to focus on them more effectively and achieve better response performance. However, when the positive units are randomly scattered throughout the knowledge base, LLMs often experience the "Lost in the Middle" (Liu et al., 2024)

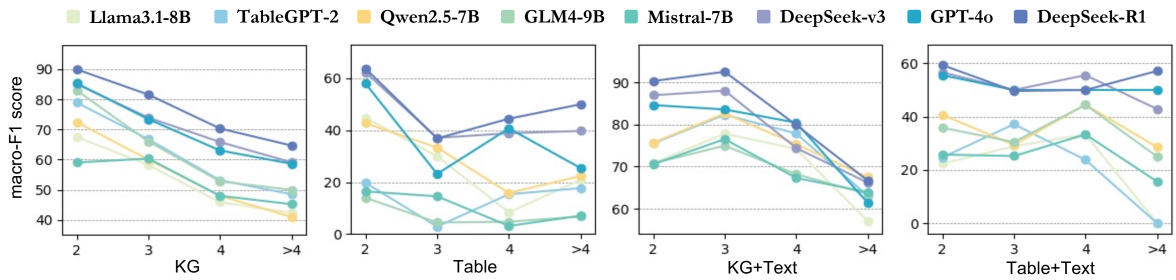

Figure 9: Information Integration results on four subsets demonstrates the variation in F1 score as the number of required positive units increases.

Model	KG	Table	KG+Text	Table+Text	Avg.
Open Source LLMs					
Llama3.1-8B	49.36	47.46	48.21	56.57	50.40
TableGPT-2	**83.69**	**70.85**	**85.13**	**93.94**	**83.40**
Qwen2.5-7B	81.55	70.17	75.90	80.81	77.11
GLM4-9B	69.96	61.69	63.59	71.72	66.74
Mistral-7B	61.37	62.71	51.28	53.03	57.10
Advanced General-Purpose LLMs					
DeepSeek-v3	78.54	69.83	58.97	69.70	69.26
GPT-4o	87.98	**73.56**	**76.92**	80.81	**79.82**
DeepSeek-R1	**91.42**	72.88	68.21	**82.32**	78.71

Table 5: Negative Rejection results on four subsets.

phenomenon, making them more likely to respond incorrectly. This suggests that for structured knowledge retrieval scenarios, **recalling positive units as early as possible can effectively enable LLMs to focus on them, thereby improving performance.**

In Fig. 7 and 8, we present the mean and standard deviation of F1 scores, and their correlation matrix across different subsets under the order insensitivity testbed. As illustrated in Fig. 8, we can observe that the order sensitivity of LLMs across 4 SK types exhibits a positive correlation, and so does their F1 performance. From the perspective of standard deviation, models that are insensitive to the order of SK are generally either weaker-performing LLMs, such as Llama3.1-8B, or exceptionally strong-performing LLMs, such as DeepSeek-R1. The former consistently exhibits weaker capabilities across various order settings, while the latter demonstrates stronger understanding and reasoning abilities, **suggesting that current LLMs are evolving towards greater robustness and less sensitive to the order of knowledge units.**

4.4 Information Integration Analysis

From the results shown in Fig. 9, it can be observed that as the number of knowledge units required increases, the overall performance of the LLMs tends to decline. This phenomenon is more pronounced in the `KG` and `KG+Text` subsets. We believe this is due to the fact that noisy knowledge units and positive knowledge units in the KG are derived from subgraph. Many noisy units share the same entities or relations as the positive units and exhibit higher semantic similarity, which more significantly impacts the LLM's understanding. In contrast, the row units of table data are relatively more semantically independent, so this downward trend is less noticeable in the `Table` subset.

In terms of understanding heterogeneous data, it is evident that as the volume of heterogeneous data increases, the performance of most LLMs declines quite substantially. Notably, in the `Table+Text` subset with >4 heterogeneous units, the advanced LLMs such as DeepSeek-R1 and GPT-4o still maintain relatively strong performance, whereas smaller LLMs like TableGPT-2 and Llama3.1-8B struggle to generate correct answers. **Thus, we consider enhancing the ability of smaller LLMs to understand heterogeneous data to be a promising research direction worthy of further exploration.**

4.5 Negative Rejection Analysis

The results in Table 5 present the rejection rates when only noisy knowledge units are provided. Overall, there is a certain positive correlation between the structured knowledge understanding performance of the LLMs and its negative rejection ability. However, we find that even DeepSeek-R1, with a negative rejection rate of 78.71%, remains vulnerable to noise interference. To our surprise, compared to Qwen2.5-7B, TableGPT-2 after fine-tuning with table-specific instructions, demonstrates stronger negative rejection ability, even surpassing GPT-4o and DeepSeek-R1. **Therefore, how to strike a balance between improving the LLM's performance and enhancing its negative rejection ability remains challenging.**

5 Conclusion

In this paper, we introduce a fine-grained structured knowledge (SK) understanding benchmark, ***SKA-Bench***, designed to provide a more comprehensive and rigorous evaluation for LLMs in understanding SK. The instances in *SKA-Bench* consist of a question, an answer, positive knowledge units, and noisy knowledge units, offering greater flexibility and scalability. Through varying the order and scale of knowledge units within the knowledge base, we construct four specialized testbeds to evaluate key capabilities: *Noise Robustness*, *Order Insensitivity*, *Information Integration*, and *Negative Rejection*. Empirical results demonstrate that even powerful LLMs like GPT-4o and DeepSeek-R1 still lack comprehensive understanding and reasoning capabilities for SK. Their performance is significantly influenced by factors such as the amount of noise, order of knowledge units, and hallucinations.

Limitations

Although *SKA-Bench* offers a more comprehensive and rigorous benchmark for evaluating structured knowledge understanding of LLMs, certain limitations warrant careful consideration, as summarized below. (1) *SKA-Bench* is limited to English only and does not yet capture the performances of LLMs in understanding structured knowledge across multiple languages. (2) Constrained by resource limitations, although our *SKA-Bench* instances have the capability to construct longer structured knowledge bases (even >64k tokens), we have not yet explored the performance of LLMs at this scale. (3) Potential biases introduced by dataset selection are inevitable. We have made effort to minimize this impact through careful dataset selection and evaluation design. For dataset selection, *SKA-Bench* adopts widely used and well-recognized benchmark datasets for each respective data type (e.g., *WebQSP*, *CWQ*, *WTQ*, *TableBench*, *StarK-Amazon*, and *HybridQA*). These datasets focus on general domains and do not require extensive domain-specific knowledge, which helps to mitigate the bias introduced by domain dependence. For evaluation design, our primary focus is on whether LLMs can comprehensively understand various types of structured knowledge, rather than investigating the "performance differences between different modalities". (4) Beyond its role as a structured knowledge understanding dataset, *SKA-Bench* can be effectively extended to various task scenarios, such as Text2Query (i.e., Text2SQL (Guo et al., 2019) and Text2SPARQL (Hu et al., 2018)), structured knowledge retrieval (Mahalingam et al., 2024), and evaluations of knowledge-augmented systems (Liang et al., 2024; Wang et al., 2025). These potential evaluation directions remain unexplored.

Ethics Statement

In this paper, we construct *SKA-Bench*, which is expanded and modified based on the existing 6 structured knowledge understanding evaluation datasets. These datasets have stated that there are no ethical concerns. Moreover, we also incorporate manual annotation and manual synthetic data verification to ensure that it does not violate any ethics.

Acknowledgments

This work is founded by National Natural Science Foundation of China (NSFC62306276/NSFCU23 B2055/NSFCU19B2027), Zhejiang Provincial Natural Science Foundation of China (No. LQ23F020017), Yongjiang Talent Introduction Programme (2022A-238-G), and Fundamental Research Funds for the Central Universities (226-2023-00138). This work was supported by Ant Group.

References

AI@Meta. 2024. Llama 3 model card.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. *Data Intelligence*, 6(3):587–620.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth*

Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 17754–17762. AAAI Press.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1026–1036. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Kenneth Huang, Bryan R. Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3697–3711. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Ritam Dutt, Sopan Khosla, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiah. 2023. Grailqa++: A challenging zero-shot benchmark for knowledge base question answering. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 897–909. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models (llms) on tabular data: Prediction, generation, and understanding - A survey. *Trans. Mach. Learn. Res.*, 2024.

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 *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18126–18134. AAAI Press.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4524–4535. Association for Computational Linguistics.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Sen Hu, Lei Zou, Jeffrey Xu Yu, Haixun Wang, and Dongyan Zhao. 2018. Answering natural language questions by subgraph matching over knowledge graphs. *IEEE Trans. Knowl. Data Eng.*, 30(5):824–837.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Jinhao Jiang, Kun Zhou, Xin Zhao, and Ji-Rong Wen. 2023b. Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024a. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024b. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. *CoRR*, abs/2410.08815.

Lei Liang, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu, Peilong Zhao, Zhongpu Bo, Jin Yang, Huaidong Xiong, Lin Yuan, Jun Xu, Zaoyang Wang, Zhiqiang Zhang, Wen Zhang, Huajun Chen, Wenguang Chen, and Jun Zhou. 2024. KAG: boosting llms in professional domains via knowledge augmented generation. *CoRR*, abs/2409.13731.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Zhiqiang Liu, Chengtao Gan, Junjie Wang, Yichi Zhang, Zhongpu Bo, Mengshu Sun, Huajun Chen, and Wen Zhang. 2025. Ontotune: Ontology-driven self-training for aligning large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 119–133.

Aakash Mahalingam, Vinesh Kumar Gande, Aman Chadha, Vinija Jain, and Divya Chaudhary. 2024. SKETCH: structured knowledge enhanced text comprehension for holistic retrieval. *CoRR*, abs/2412.15443.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1470–1480. The Association for Computer Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10):197:1–197:45.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. 2024. Tablegpt2: A large multimodal model with tabular data integration. *Preprint*, arXiv:2411.02059.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 641–651. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jinyu Wang, Jingjing Fu, Rui Wang, Lei Song, and Jiang Bian. 2025. PIKE-RAG: specialized knowledge and rationale augmented generation. *CoRR*, abs/2501.11551.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N. Ioannidis, Karthik Subbian, James Y. Zou, and Jure Leskovec. 2024. Stark: Benchmarking LLM retrieval on textual and relational knowledge bases. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. 2025. Tablebench: A comprehensive and complex benchmark for table question answering. pages 25497–25506.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang.

2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3277–3287. Association for Computational Linguistics.

A Original Datasets Details

We provide a brief description of all the original structured knowledge understanding datasets we used and licenses below:

- ***WebQSP*** (Yih et al., 2016). WEBQUESTIONSSP (*WebQSP*) is a semantic parse-based KBQA dataset with 4,737 questions coupled with SPARQL queries for KB question answering. The answers can be extracted through executing SPARQL queries on Freebase. The dataset is released under the Microsoft Research Data License Agreement.
- ***CWQ*** (Talmor and Berant, 2018). COMPLEXWEBQUESTIONS (*CWQ*) is created on top of *WebQSP* dataset with the intention of generating more complex (by incorporating compositions, conjunctions, superlatives or comparatives) questions in natural language. It consists of 34,689 examples, divided into 27,734 train, 3,480 dev, 3,475 test. And test set in original *CWQ* dataset does not contain "answer". The whole software is licensed under the full GPL v2+.
- ***WTQ*** (Pasupat and Liang, 2015). WIKITABLEQUESTIONS (*WTQ*) is a widely used table question answering (TableQA) dataset of 22,033 complex questions with average 2.14 hop on Wikipedia tables. The dataset is released under the Apache-2.0 license.
- ***TableBench*** (Wu et al., 2025). *TableBench* is a comprehensive and complex benchmark, including 886 samples in 18 fields within four major categories of TableQA capabilities. The tables in *TableBench* have an average of 6.68 columns and 16.71 rows, and the average reasoning steps of questions is 6.26. The dataset is released under the Apache-2.0 license.
- ***STaRK*** (Wu et al., 2024). *STaRK* is a large-scale semi-structure retrieval benchmark on textual and relational knowledge bases, covering three domains. It consists of 263 human-generated questions and 33,627 synthesized questions. And this dataset is released under the MIT license.
- ***HybridQA*** (Chen et al., 2020b). *HybridQA* is a question answering dataset based on heterogeneous knowledge, and each question is aligned with a Wikipedia table and multiple free-form corpora linked with the entities in the table. The questions are collected from crowd-workers, and designed to aggregate both table and text information, which means the lack of either form would render the question unanswerable. The dataset is released under the MIT license.

Type	1-hop	2-hop	3-hop	≥4-hop
KG subset	7.59%	22.07%	16.26%	54.09%
KG+Text subset	7.04%	10.91%	9.76%	72.29%

Table 6: Hop count distributions in KG subset and KG+Text subset.

B Effects of Noise Distribution in KG-related Subsets

Due to the structural complexity of knowledge graphs, the degree of noise can vary across different triples. We hypothesize that hop count is closely related to semantic similarity, and this relationship may affect the noise distribution, thereby impacting model performance. To investigate this, we conduct an in-depth analysis of the hop count distribution in the existing KG subsets and KG+Text subsets, and provide experimental results to support this hypothesis. As shown in Table 6, we present the detailed hop count distributions for the KG and KG+Text subsets. It is worth noting that we report the actual distributions in *SKA-Bench* rather than the original datasets, since some subgraphs were randomly truncated during the graph extraction process.

Model	KG				KG+Text			
	Hard Noise	Mixed Noise	Easy Noise	Std.	Hard Noise	Mixed Noise	Easy Noise	Std.
LLaMA3.1-8B	55.12	58.19	60.23	2.10	65.68	67.02	66.96	0.62
Qwen2.5-7B	58.36	60.00	67.82	4.13	74.96	76.51	77.15	0.92
DeepSeek-v3	71.09	**73.93**	74.28	1.43	**77.01**	77.12	77.65	**0.28**
GPT-4o	**72.10**	73.42	**74.82**	**1.11**	76.58	**77.38**	**77.63**	0.45

Table 7: Performance of four representative LLMs under different noise intensities on KG and KG+Text subsets.

Thanks to the high flexibility of *SKA-Bench*, we further present results under different noise intensities to more thoroughly evaluate the noise robustness of LLMs. In our experimental setup, the context size is consistently controlled at 4k, and knowledge units from the knowledge graph are arranged in a random order. Noise is categorized as "Hard Noise" ($\leq$2 hops), "Mixed Noise" (same as in main text), and "Easy Noise" ($\geq$4 hops). We report the performance of four representative LLMs, as shown in Table 7. The experimental results show that the variation in model performance is significantly correlated with the intensity of noise. Comparatively, stronger LLMs demonstrate greater robustness to noise, and this conclusion also holds when the noise level is expanded horizontally (as in the noise robustness testbed of the main text).

C Dataset Construction Details

The annotation guideline for "Iterative Positive Units Annotation" is shown in the Fig. 10.

Moreover, we have presented specific examples of the part where LLMs are involved in the entire dataset construction process. Check satisfaction by LLM in Iterative Positive Units Annotation stage is shown in Fig. 11. Noisy Synthesis process is shown in Fig. 12. And "condition" of positive knowledge units summarizing and check satisfaction by LLMs in Noisy Units Construction stage are shown in Fig. 13 and Fig. 14.

D Evaluation Prompt Template

Fig. 15, 16, 17, 18 show QA prompt templates for four subsets in *Noise Robustness* testbed, *Order Insensitivity* testbed, and *Information Integration* testbed. Fig. 19, 20, 21, 22 show prompt templates of *negative rejection* testbed for four subsets.

Annotation Guidelines

With the improvement of the structured knowledge understanding ability of large language models, the existing structured knowledge understanding evaluations are difficult to fully diagnose the shortcomings of LLMs. Therefore, we invite you to annotate the **positive knowledge units** from the whole knowledge base of the following structured knowledge augmented QA, thereby obtaining a more complex and comprehensive structured knowledge evaluation dataset. Our annotation instructions are as follows:

(1) if the answer is wrong, delete the sample directly;
(2) if the question involves multiple answers, all positive units require to obtain the answers should be annotated;
(3) for the Table subset and Table+Text subset, if the question needs to perform numerical analysis on the entire table, the corresponding SKA-QA pairs should either be removed or the question should be modified;
(4) if the tables in the Table subset and Table+Text subset are order-dependent (i.e., modifying the row order would result in semantic errors in the table), this sample should be removed;
(5) for the KG+Text subset and Table+Text subset, if question only utilizes one type of knowledge source, the question should be modified or removed.


```
"id": "TableBench-3ba617b11797d522d5c51bdb6d6cd313",
"question": "What is the total population of all regions in China where the percentage in Manchu population is greater than 5%?",
"answer": [
    "3123625869"
],
"positive_rows": [],
"header": "|region|total population|manchu|percentage in manchu population|regional percentage of population|",
"original_rows": [
    "|total|1335110869|10410585|100|0.77|",
    "|total (in all 31 provincial regions)|1332810869|10387958|99.83|0.78|",
    "|northeast|109513129|6951280|66.77|6.35|",
    "|north|164823663|3002873|28.84|1.82|",
    "|east|392862229|122861|1.18|0.03|",
    "|south central|375984133|120424|1.16|0.03|",
    "|northwest|96646530|82135|0.79|0.08|",
    "|southwest|192981185|57785|0.56|0.03|",
    "|liaoning|43746323|5336895|51.26|12.2|",
    "|hebei|71854210|2118711|20.35|2.95|",
    "|jilin|27452815|866365|8.32|3.16|",
    "|heilongjiang|38313991|748020|7.19|1.95|",
    "|inner mongolia|24706291|452765|4.35|2.14|",
    "|beijing|19612368|336032|3.23|1.71|",
    "|tianjin|12938693|83624|0.80|0.65|",
    "|henan|94029939|55493|0.53|0.06|",
    "|shandong|95792719|46521|0.45|0.05|",
    "|guangdong|104320459|29557|0.28|0.03|",
    "|shanghai|23019196|25165|0.24|0.11|",
    "|ningxia|6301350|24902|0.24|0.4|",
    "|guizhou|34748556|23086|0.22|0.07|",
    "|xinjiang|21815815|18707|0.18|0.09|",
    "|jiangsu|78660941|18074|0.17|0.02|",
    "|shaanxi|37327379|16291|0.16|0.04|",
```


Figure 10: The annotation guidelines for annotators.

Question:
Which nation has the Alta Verapaz Department and is in Central America?
Answer:
Guatemala
The above are questions and above all answers. Please judge whether the following triples can deduce answers to the above questions. If you can get partial answers, reply me "1" directly; If you can get all the answers, please reply me "2" directly; If you can't get any result, please reply me "0" directly.
Triples:
(Guatemala, location.country.administrative_divisions, Alta Verapaz Department),
(Central America, location.location.contains, Guatemala),
(Guatemala, common.topic.notable_types, Country)

Figure 11: The prompt for checking Positive Units.

Table:
Date	Site	Winning team	Winning team score	Losing team	Losing team score	Series
September 6, 1980	Fort Collins	Colorado State	21	Air Force	9	AFA 11–7–1
October 3, 1981	Colorado Springs	Air Force	28	Colorado State	14	AFA 12–7–1
October 16, 1982	Colorado Springs	Colorado State	21	Air Force	11	AFA 12–8–1
September 26, 1987	Fort Collins	Air Force	27	Colorado State	19	AFA 17–8–1
September 3, 1988	Fort Collins	Air Force	29	Colorado State	23	AFA 18–8–1
October 17, 1992	Colorado Springs	Colorado State	32	Air Force	28	AFA 20–10–1
September 11, 1993	Fort Collins	Colorado State	8	Air Force	5	AFA 20–11–1
September 3, 1994	Colorado Springs	Colorado State	34	Air Force	21	AFA 20–12–1
September 16, 1995	Colorado Springs	Colorado State	27	Air Force	20	AFA 20–13–1
November 2, 1996	Colorado Springs	Colorado State	42	Air Force	41	AFA 20–14–1
September 20, 1997	Fort Collins	Air Force	24	Colorado State	0	AFA 21–14–1
September 17, 1998	Colorado Springs	Air Force	30	Colorado State	27	AFA 22–14–1
November 18, 1999	Fort Collins	Colorado State	41	Air Force	21	AFA 22–15–1
November 11, 2000	Colorado Springs	Air Force	44	Colorado State	40	AFA 23–15–1
November 8, 2001	Fort Collins	Colorado State	28	Air Force	21	AFA 23–16–1
October 31, 2002	Colorado Springs	Colorado State	31	Air Force	12	AFA 23–17–1
October 16, 2003	Fort Collins	Colorado State	30	Air Force	20	AFA 23–18–1
November 20, 2004	Colorado Springs	Air Force	47	Colorado State	17	AFA 24–18–1
September 29, 2005	Fort Collins	Colorado State	41	Air Force	23	AFA 24–19–1
Task Description: According to the above table, we have the following question and answer.
Question: which date the colorado state team scored no points?
Answer: September 20, 1997
Your task is to generate 20 noisy rows for the table. You need to make sure that you don't change the answer to the current question after adding noise rows to the table. Your output noise rows must not duplicate the existing table, and the table format should be the same as the original table. Note that your output does not contain the original table rows.

Figure 12: The prompt for Noisy Units synthesis.

Question:
which date the colorado state team scored no points?
Answer:
September 20, 1997
Positive Units:
|Date|Site|Winning team|Winning team score|Losing team|Losing team score|Series|
|September 20, 1997|Fort Collins|Air Force|24|Colorado State|0|AFA 21–14–1|
Task Description: The above is a KBQA question, the answer, and the positive knowledge unit necessary to answer it. Please help me summarize what "conditions" the noise knowledge unit that cannot be used to answer this question needs to meet in the last line.

output: Conditions: The knowledge unit does not involve Colorado State as the losing team with a score of 0.

Figure 13: The prompt for "contidition" summarizing.

Question:
which date the colorado state team scored no points?
Answer:
September 20, 1997
Noisy Units:
Date	Site	Winning team	Winning team score	Losing team	Losing team score	Series
November 15, 1984	Boulder	Colorado State	40	Wyoming	25	CSU 5–3
October 14, 1992	Fort Collins	Utah	28	Colorado State	19	Utah 8–1
October 21, 2007	Colorado Springs	Air Force	35	Wyoming	10	AFA 16–3,
September 12, 1996	Boulder	Colorado	28	Minnesota	17	CU 12–2
October 23, 1982	Albuquerque	New Mexico	30	Air Force	24	NM 6–5
November 4, 1998	Tuscaloosa	Alabama	37	LSU	34	UA 15–7
September 10, 2001	Denver	Raiders	27	Denver	24	Raiders 3–6
October 1, 2005	Boulder	Colorado	38	Kansas	21	CU 9–0
November 18, 1995	Fort Collins	Colorado State	45	BYU	29	CSU 10–5
October 27, 1988	Colorado Springs	Air Force	41	Navy	19	AFA 17–4
October 14, 1995	Denver	Seattle	28	Denver	17	Seahawks 1–0
November 23, 2006	Fort Collins	Fort Collins	31	San Diego	30	FC 1–0
September 5, 1989	Boulder	Texas	17	California	9	Texas 1–0
October 15, 1993	Colorado Springs	Arizona	41	New Mexico	10	AZ 2–0
November 1, 2007	Denver	Denver	23	Chiefs	17	Broncos 7–0
December 8, 1984	Boulder	South Dakota	35	Boston College	25	SD 1–0
September 29, 1999	Fort Collins	Utah	29	Air Force	22	Utah 2–1
October 2, 2002	Tuscaloosa	Alabama	28	Southern Miss	21	UA 3–0
November 20, 2010	Colorado Springs	Texas Tech	42	Colorado State	7	TTU 1–0
September 6, 1994	Colorado Springs	Notre Dame	24	Kansas	22	ND 2–0
Positive Unit Condition:
Conditions: The knowledge unit does not involve Colorado State as the losing team with a score of 0.
Task Description: The above are KBQA question and corresponding answer. Please judge whether the Noisy knowledge units satisfy the "positive unit condition", thereby deducing answer to the above question. If you can get partial answers, reply me "1" directly; If you can get all the answers, please reply me "2" directly; If you can't get any result, please reply me "0" directly.

Figure 14: The prompt for checking Noisy Units.

Triples:
(Guatemala, location.location.containedby, North America)
(Guatemala, book.book_subject.works, Tree Girl)
(Denmark, location.location.containedby, Scandinavia)
(German state, type.type.domain, Location)
(The Jaguar Smile, book.book.editions, The Jaguar Smile)
(Hondo River, location.location.containedby, North America)
(Bunnik Tours, business.brand.owner_s, m.012m0fnn) **… …**
Task Description: Based on the triples provided above, please answer the following questions.
Question: What language is spoken in the location that appointed Michelle Bachelet to a governmental position speak?
Return the final result as JSON in the format {"answer": <YOUR ANSWER LIST>} in the last line.

Figure 15: The prompt for `KG` subset in QA task.

Table:
Iteration	Year	Dates	Location	Theme
1st	1972	6 May-20 May	Suva, Fiji	"Preserving culture"
2nd	1976	6 March-13 March	Rotorua, New Zealand	"Sharing culture"
3rd	1980	30 June-12 July	Port Moresby, Papua New Guinea	"Pacific awareness"
4th	1985	29 June-15 July	Tahiti, French Polynesia	"My Pacific"
5th	1988	14 August-24 August	Townsville, Australia	"Cultural interchange"
6th	1992	16 October-27 October	Rarotonga, Cook Islands	"Seafaring heritage"
7th	1996	8 September-23 September	Apia, Sāmoa	"Unveiling treasures"
Task Description: Please look at the table, and then answer the following questions.
Question: what is the number of themes that refer to "culture"?
Return the final result as JSON in the format {"answer": <YOUR ANSWER LIST>} in the last line.

Figure 16: The prompt for `Table` subset in QA task.

Triples:
(PARP1, expression_present, cerebellar cortex)
(VCP, ppi, HSPA5)
(Elevated hepatic transaminase, associated_with, SOCS1)
(PSMC5, expression_present, nasal cavity mucosa) **… …**
Texts:
- name: toxic epidermal necrolysis\n- type: disease - source: MONDO - details: - mondo_name: toxic epidermal necrolysis\n - mondo_definition: Toxic epidermal necrolysis (TEN) is an acute and severe skin disease with clinical and histological features characterized by the destruction and detachment of the skin epithelium and mucous membranes. - umls_description: A systemic, serious, and life-threatening disorder characterized by erythematous and necrotic lesions in the skin and mucous membranes that are associated with bullous detachment of the epidermis. **… …**
Task Description: Based on the triples and texts provided above, please answer the specific product for following questions.
Question: I have nail dystrophy and chemosis. What skin disease might I have?
Return the final result as JSON in the format {"answer": <YOUR ANSWER LIST>} in the last line.

Figure 17: The prompt for `KG+Text` subset in QA task.

Table:
Name	Years	Apps	Goals	Position
Billy Bassett	1886-99	311	77	Outside right
Jesse Pennington	1903-22	496	0	Left back
W. G. Richardson	1929-45	354	228	Centre forward
Ray Barlow	1944-60	482	48	Left-half	**… …**
Texts:
Bryan Robson: Bryan Robson OBE (born 11 January 1957) is an English football manager and former player. Born in Chester-le-Street, County Durham, he began his career with West Bromwich Albion in 1972 before moving to Manchester United in 1981, where he became the longest serving captain in the club's history and won two Premier League winners' medals, three FA Cups, two FA Charity Shields and a European Cup Winners' Cup. **… …**
Task Description: Based on the table and texts provided above, please answer the specific product for following questions.
Question: What are the goals of the athlete who initiated his management career as a player-manager with Middlesbrough in 1994?
Return the final result as JSON in the format {"answer": <YOUR ANSWER LIST>} in the last line.

Figure 18: The prompt for `Table+Text` subset in QA task.

Triples:
(Guatemala, location.location.containedby, North America)
(Guatemala, book.book_subject.works, Tree Girl)
(Denmark, location.location.containedby, Scandinavia)
(German state, type.type.domain, Location)
(The Jaguar Smile, book.book.editions, The Jaguar Smile)
(Hondo River, location.location.containedby, North America)
(Bunnik Tours, business.brand.owner_s, m.012m0fnn) **… …**
Task Description: Based on the triples provided above, please judge whether the following questions can be answered.
Question: what language is spoken in the location that appointed Michelle Bachelet to a governmental position speak?
Return the final result as JSON in the format {"answer": "yes"} or {"answer": "no"} in the last line.

Figure 19: The prompt for `KG` subset in "*negative rejection*" testbed.

Table:
Iteration	Year	Dates	Location	Theme
1st	1972	6 May-20 May	Suva, Fiji	"Preserving culture"
2nd	1976	6 March-13 March	Rotorua, New Zealand	"Sharing culture"
3rd	1980	30 June-12 July	Port Moresby, Papua New Guinea	"Pacific awareness"
4th	1985	29 June-15 July	Tahiti, French Polynesia	"My Pacific"
5th	1988	14 August-24 August	Townsville, Australia	"Cultural interchange"
6th	1992	16 October-27 October	Rarotonga, Cook Islands	"Seafaring heritage"
7th	1996	8 September-23 September	Apia, Sāmoa	"Unveiling treasures"
Task Description: Please look at the table, and then judge whether the following questions can be answered.
Question: what is the number of themes that refer to "culture"?
Return the final result as JSON in the format {"answer": "yes"} or {"answer": "no"} in the last line.

Figure 20: The prompt for `Table` subset in "*negative rejection*" testbed.

Triples:
(PARP1, expression_present, cerebellar cortex)
(VCP, ppi, HSPA5)
(Elevated hepatic transaminase, associated_with, SOCS1)
(PSMC5, expression_present, nasal cavity mucosa) **… …**
Texts:
- name: toxic epidermal necrolysis\n- type: disease - source: MONDO - details: - mondo_name: toxic epidermal necrolysis\n - mondo_definition: Toxic epidermal necrolysis (TEN) is an acute and severe skin disease with clinical and histological features characterized by the destruction and detachment of the skin epithelium and mucous membranes. - umls_description: A systemic, serious, and life-threatening disorder characterized by erythematous and necrotic lesions in the skin and mucous membranes that are associated with bullous detachment of the epidermis. **… …**
Task Description: Based on the triples and texts provided above, please judge whether the following questions can be answered.
Question: I have nail dystrophy and chemosis. What skin disease might I have?
Return the final result as JSON in the format {"answer": "yes"} or {"answer": "no"} in the last line.

Figure 21: The prompt for `KG+Text` subset in "*negative rejection*" testbed.

Table:
Name	Years	Apps	Goals	Position
Billy Bassett	1886-99	311	77	Outside right
Jesse Pennington	1903-22	496	0	Left back
W. G. Richardson	1929-45	354	228	Centre forward
Ray Barlow	1944-60	482	48	Left-half	**… …**
Texts:
Bryan Robson: Bryan Robson OBE (born 11 January 1957) is an English football manager and former player. Born in Chester-le-Street, County Durham, he began his career with West Bromwich Albion in 1972 before moving to Manchester United in 1981, where he became the longest serving captain in the club's history and won two Premier League winners' medals, three FA Cups, two FA Charity Shields and a European Cup Winners' Cup. **… …**
Task Description: Based on the table and texts provided above, please judge whether the following questions can be answered.
Question: What are the goals of the athlete who initiated his management career as a player-manager with Middlesbrough in 1994?
Return the final result as JSON in the format {"answer": "yes"} or {"answer": "no"} in the last line.

Figure 22: The prompt for `KG+Text` subset in "*negative rejection*" testbed.