

MT-RAIG: Novel Benchmark and Evaluation Framework for Retrieval-Augmented Insight Generation over Multiple Tables

Kwangwook Seo* Donguk Kwon* Dongha Lee†

Yonsei University - DLI Lab

{tommy2130, donguk.kwon, donalee}@yonsei.ac.kr

Abstract

Recent advancements in table-based reasoning have expanded beyond factoid-level QA to address insight-level tasks, where systems should synthesize implicit knowledge in the table to provide explainable analyses. Although effective, existing studies remain confined to scenarios where a single gold table is given alongside the user query, failing to address cases where users seek comprehensive insights from multiple unknown tables. To bridge these gaps, we propose **MT-RAIG BENCH**, designed to evaluate systems on **Retrieval-Augmented Insight Generation over Multiple-Tables**. Additionally, to tackle the suboptimality of existing automatic evaluation methods in the table domain, we further introduce a fine-grained evaluation framework **MT-RAIG EVAL**, which achieves better alignment with human quality judgments on the generated insights. We conduct extensive experiments and reveal that even frontier LLMs still struggle with complex multi-table reasoning, establishing our MT-RAIG BENCH as a challenging testbed for future research¹.

1 Introduction

Tables are ubiquitous in the real-world data and constitute a significant portion of the information available on the web and databases. While their structured nature efficiently encapsulates diverse information, it also poses challenges for developing robust table understanding systems (Pasupat and Liang, 2015a; Chen et al., 2020; Tang et al., 2024; Seo et al., 2024). This inherent complexity of table data has led to a persistent demand for the systems capable of faithfully interpreting table content and presenting it to users in a human-readable format.

In response to these needs, existing table-based question answering (TQA) works (Pasupat and Liang, 2015a; Nan et al., 2021; Wang et al., 2024b)

* Equal contribution

† Corresponding author

¹<https://kwondu.github.io/mt-raig>

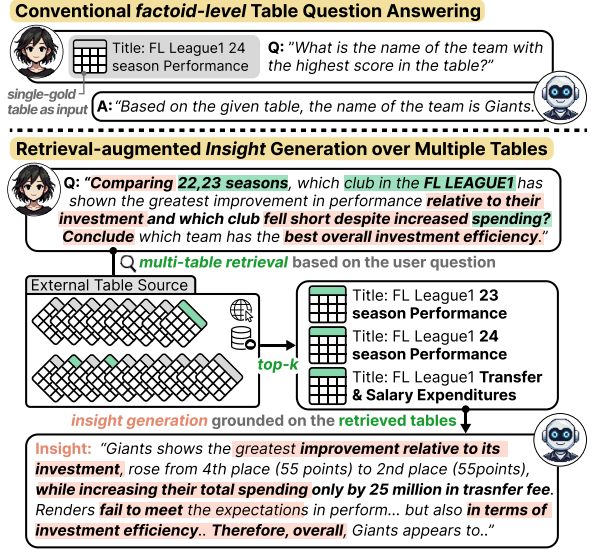


Figure 1: Comparison between MT-RAIG and conventional table question answering task.

have predominantly focused on extracting explicit facts from a given table by developing systems that can follow the detailed instructions outlined in user queries. However, these *factoid-level* queries often constrain the systems' operation to functioning as an extractive executor, reducing its role to retrieving only a fraction of values explicitly presented in the table. This narrow focus prevents the systems from comprehensively analyzing the table's full context, which is essential for uncovering implicit information embedded within the table.

Recently, some studies (Moosavi et al., 2021; Zhao et al., 2023a; Seo et al., 2024) have begun to explore more practical user information needs by moving beyond this traditional *factoid-level* scenario. These studies address *insight-level* table reasoning, wherein the user requires not just simple fact retrieval but also deeper insight mining from the table. This shift has led to the development of advanced agents designed to deliver explainable analyses and meaningful data insights, accommodating scenarios where users seek comprehensive interpretations and synthesized knowledge.

Dataset	#Test Examples	#Unique Tables	#Words /Output	#Tables /Example	Retrieval Availability	Reasoning Depth
<i>Text-to-SQL generation</i>						
WikiSQL (Zhong et al., 2018)	15,878	26,531	6.13	1.00	✗	factoid-level
SPIDER (Yu et al., 2018)	1,034	876	18.37	1.58	✗	factoid-level
<i>Table-to-text generation</i>						
ToTTo (Parikh et al., 2020)	7,700	83,141	17.37	1.00	✗	factoid-level
InsTaSumm (Seo et al., 2024)	440	2,494	161.90	1.00	✗	insight-level
<i>Table question-answering</i>						
WTQ (Pasupat and Liang, 2015b)	9,473	2,108	1.76	1.00	✗	factoid-level
FeTaQA (Nan et al., 2021)	2,003	10,330	23.30	1.00	✗	factoid-level
Open-WikiTable (Kweon et al., 2023)	6,602	24,680	1.90	1.00	✓	factoid-level
QTSumm (Zhao et al., 2023a)	1,078	2,934	67.76	1.00	✗	insight-level
MT-RAIG BENCH (ours)	18,532	19,563	189.87	2.88	✓	insight-level

Table 1: Comparison with existing benchmarks for table-related tasks.

Despite these advancements in table-based reasoning, two critical limitations remain unaddressed. **(1)** Existing approaches typically operate in a **closed-domain setting**, where a pre-defined gold table is provided alongside the query at test time. While it simplifies the testbed for evaluating the system’s reasoning ability, it also introduces a significant drawback. In particular, requiring users to manually craft table inputs for every query is both costly and unrealistic, as users often lack prior knowledge about which specific tables are relevant to their needs. **(2)** Most prior works focus on **single-table tasks**, assuming all source information for the reasoning is contained within a single table. However, considering that users’ information needs may require comprehensive insights spanning multiple aspects across tables, this scenario falls short in robustly handling diverse user needs.

To bridge these gaps, we propose **MT-RAIG BENCH**, aiming to measure the system’s ability on **Retrieval-Augmented Insight Generation over Multiple-Tables**. To enable more practical applications of the table-based system, our MT-RAIG BENCH requires the system to retrieve multiple evidence tables based on the input query and integrate information across them to generate insightful response. Compared to existing benchmarks, this task introduces new challenges by requiring the systems not only to retrieve query-relevant tables but also to faithfully extract evidences scattered among irrelevant tables and comprehensively aggregate these evidences to derive meaningful insight.

Although MT-RAIG BENCH can serve as a promising testbed for insight-level table reasoning, reliably evaluating long-form outputs in table-based tasks remains a longstanding challenge in the

field (Zhao et al., 2024). While recent studies (Zhao et al., 2023b; Wang et al., 2024a) employ table-specific metrics that assess the quality of generated outputs beyond the surface-level matching, they still struggle to evaluate the quality of the insight on the MT-RAIG task. Such limitation arise from their reliance on coarse-grained analyses, which fall short in detecting the finer distinctions needed to check both the output’s groundness on multi-tables and completeness on a multi-hop query.

In light of these challenges, we propose a novel decomposition-based evaluation framework **MT-RAIG EVAL**. To reliably assess the finer quality of long-form insights in MT-RAIG BENCH, MT-RAIG EVAL performs: (1) *table-aware insight decomposition* to verify the explicit grounding between the fact-entangled insight and retrieved tables, (2) *question-aware insight decomposition* to check whether the key steps required to address the question are completely followed in the insight. Our meta-evaluation validates that MT-RAIG EVAL outperforms conventional metrics in aligning with human judgments, reliably assessing both faithfulness and completeness of generated insights. We summarize our contributions as follows:

- We propose MT-RAIG BENCH, the first large-scale benchmark for retrieval-augmented insight generation over multiple tables.
- We introduce MT-RAIG EVAL, a novel automated evaluation framework that assesses the fine-grained quality of the generated insights.
- We evaluate various LLMs and SOTA methods, revealing that existing models still struggle with multi-table reasoning, establishing our benchmark as a challenging testbed for future research.

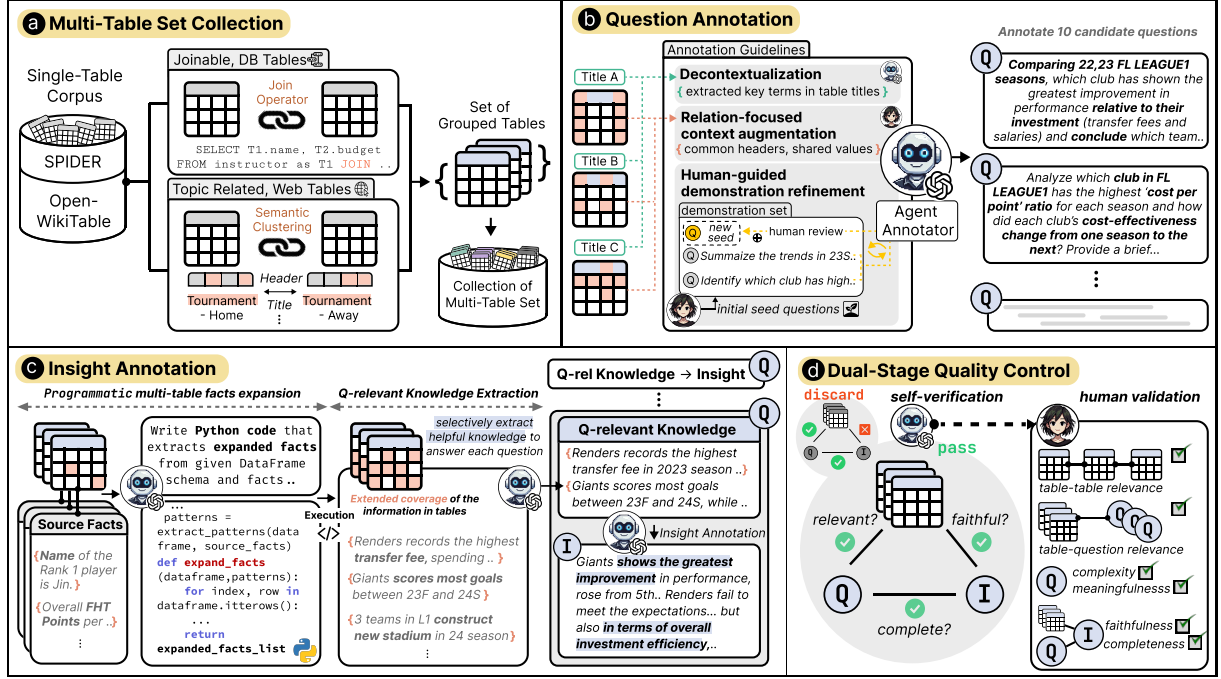


Figure 2: Overview of our MT-RAIG BENCH construction process.

2 MT-RAIG BENCH

2.1 Task Formulation

In line with the recent advancements of retrieval-augmented generation (RAG) in diverse domains (Ram et al., 2023; Shi et al., 2024), MT-RAIG task consists of two main steps: (1) *table retrieval*, (2) *insight generation*. Formally, given a natural language question q and an external data-source with tables T , the task is to first retrieve a set of evidence tables $\hat{T}_q \subset T$ to approximate the gold table set T_q relevant to q . Subsequently, an insight i is generated by grounding it on the retrieved \hat{T}_q . These two steps can be formulated as follows:

$$\hat{T}_q = \text{Ret}(q, T), \quad i = \text{Gen}(q, \hat{T}_q) \quad (1)$$

2.2 Benchmark Construction

We provide detailed information on the construction of the MT-RAIG BENCH. Specifically, we describe the process of creating a machine-annotated dataset, building on efforts from previous works (Tang and Yang, 2024; Wei et al., 2024; Ni et al., 2024; Kim et al., 2024c; Yao et al., 2024) while leveraging several notable advantages over fully human-annotated benchmark. These benefits include: (1) **scalability**, where large-scale expansion is possible with significantly reduced human labor, and (2) **consistency** in labeling standards, a crucial aspect that can be more challenging to maintain when multiple human annotators are involved.

Additionally, we incorporate human quality checks as critical review points for the automatically generated data, balancing the efficiency of automated annotation with the reliability of human annotation to ensure the high-quality of MT-RAIG BENCH. We provide all the detailed prompts in Appendix A.1 and benchmark examples in Appendix E.

2.2.1 Multi-Table Set Collection

The goal of this step is to extend the existing single-table corpus into a collection of multi-table sets. To group individual tables in the data source into coherent sets that are relevant to real user queries, we classify scenarios where multiple tables are combined into two categories: *joinable* and *topic-related*. Joinable tables can be directly linked through common key columns, following similar concept to standard text-to-SQL tasks. To construct such sets, we use SPIDER (Yu et al., 2018) as a source dataset, leveraging foreign key connections to identify joinable tables from single-table corpus.

While joinable tables offer a straightforward approach to combining individual tables, real-world table sets often present more complex relationships, where tables cannot be directly joined via a common key column. Instead, these tables are loosely connected by shared topics or contextual relevance. To effectively group these tables, we leverage table titles and headers as semantic indicators to cluster topically related tables. We source these tables

from Open-WikiTable (Kweon et al., 2023), a collection of tables extracted from Wikipedia, where rich metadata provides the necessary cues for collecting topically coherent multi-table sets.

2.2.2 Question Annotation

Unlike existing factoid-level questions, our approach aims to annotate questions that seek comprehensive insights across tables. Following previous work (Kim et al., 2024c), we employ GPT-4o mini as an agent annotator, and generate 10 distinct questions for each table set, designed to capture diverse relational aspects among the tables. To strike a robust middle ground between the reliability of human annotation and the versatility of LLMs, we adopt a *human-in-the-loop* process that faithfully guides the question annotation while preserving the flexibility of LLMs. Specifically, the question annotation process incorporates three key methods:

(1) Decontextualization: The goal of decontextualization (Choi et al., 2021; Kweon et al., 2023) is to enhance the clarity of how each question links back to the relevant tables. By explicitly including keywords derived from the table titles into the questions, this step ensures that the semantic alignment between the question and the tables become apparent. To achieve this, the agent annotator first extracts key terms from table titles and incorporate them directly into the phrasing of the questions.

(2) Relation-focused context augmentation: This step aims to highlight relationships between tables by enriching the questions with contextual information shared across the tables. We find that naively prompting the agent annotator to generate questions for multiple sources often leads to mere concatenations of information separately extracted from each table, resulting in questions that fail to reflect the connections between the tables. To address this issue, we first manually identify common attributes and shared data points across multiple tables and let the agent annotator incorporate these overlapping values into the question generation to ensure that critical relational cues between the tables are preserved within the generated questions.

(3) Human-guided demonstration refinement: To guide the agent in generating questions that are both diverse and aligned with our annotation objectives, we adopt an iterative workflow that combines agent-based automation with human feedback. We begin by categorizing questions into distinct types as shown in Table 2, and annotate initial human-written questions for each type. These seed ques-

tions serve as initial demonstrations for agent annotator to generate new questions. Then, if human reviews confirm a novel pattern from generated questions, it is incorporated back into the seed set. Through this iterative human-guided refinement, we build a final demonstration set to help the agent generate well-grounded questions for each type while covering diverse multi-hop relationships.

Properties	Value
# of Test examples	18,532
# of Words / insight	189.87
# of Unique tables	19,563
# of Unique table set	5,418
# of Gold tables / example	2.88
# of Rows / table	10.54
# of Columns / table	6.04
Question Types	Size
Analysis & Summary (A&S)	1,886 (10%)
Comparison & Relationship (C&R)	4,035 (22%)
Performance & Outcome (P&O)	10,141 (55%)
Trend & Pattern (T&P)	2,470 (13%)

Table 2: Basic statistics of MT-RAIG BENCH

2.2.3 Insight Annotation

Programmatic multi-table facts expansion To enhance the agent annotators’ understanding of structured data, we augment multi-table set with natural language (NL) facts that provide additional context for each table set. These facts are initialized from human-annotated sources in existing datasets (SPIDER, Open-WikiTable), which are accurate but focus narrowly on specific table portions, lacking comprehensive coverage of the full table context. To address this, we employ a programmatic fact expansion process to cover a broader range of information within the tables. Specifically, given the source tables and initial NL facts, the agent annotator generates a Python function, `expand_facts`. This function is then executed on the tables to systematically extract enriched facts while ensuring faithfulness to the source table.

Question-relevant knowledge extraction Based on the enriched facts obtained in the previous step, the next step involves annotating insights that serve as a comprehensive answer to the given question. However, even when providing the agent annotator with enriched NL contexts derived from fact expansion, such contexts might introduce noise and hinder the insight annotation, as not all information helps address each question. Therefore, we filter out irrelevant content for each question and selectively extract only the relevant facts to pro-

vide the agent annotator with a condensed form of knowledge. This knowledge is then fed to the agent annotators to generate insights for each table set.

2.2.4 Dual-Stage Quality Control

For the last step of our benchmark construction, we adopt dual-stage quality assurance process that combines human and agent verification. First, following [Tang and Yang \(2024\)](#), we utilize an agent annotator as a self-verifier, ensuring that each multi-table set, question, and insight triple satisfies strict criteria for *relevance*, *faithfulness*, and *completeness*—discarding any triple that fails at least one of these standards. Second, we conduct a human validation, wherein the machine-verified samples are manually reviewed according to the criteria in Table 3, thereby confirming the accuracy and coherence of the agent-based annotations. Table 3 summarizes this result, showing high degree of agreement among human evaluators, demonstrating the high-quality of our benchmark. We provide the case study, statistics of self-verification, and detailed process of quality control in Appendix A.2.

Data Quality	%S _{≥4}	Agree	Kappa
Inter-Table Relevance	97.33	0.83	0.78
Table-Question Relevance	96.00	0.87	0.84
Question Complexity	98.67	0.87	0.83
Question Meaningfulness	98.00	0.89	0.86
Question-Insight Completeness	94.00	0.85	0.81
Table-Insight Faithfulness	89.67	0.86	0.82

Table 3: Human evaluation over 300 samples of MT-RAIG BENCH. Two evaluators are asked to rate the samples in 1-5 Likert Scale. We report percentage of samples that have average score above 4, percent of evaluator agreement and Cohen’s Kappa with 95% CI.

3 MT-RAIG EVAL

Automatic evaluation on the quality of long-form output in table-based tasks remains a long-standing challenge that often does not align well with human evaluations ([Zhao et al., 2024](#); [Wang et al., 2024a](#); [Seo et al., 2024](#)). This issue is even more pronounced in the MT-RAIG task, as it is challenging to check (1) the explicit grounding between fact-entangled insight and multiple tables, (2) whether the key steps required to address the multi-hop questions are completely followed in insight. We posit that these challenges arise as existing automatic evaluation methods (*e.g.*, BLEU ([Papineni et al., 2002](#)), TAPAS-Acc ([Liu et al., 2022](#)), and G-Eval ([Liu et al., 2023a](#))) often analyze the output

in a coarse-grained manner. To tackle this challenge, we propose a novel **decomposition-based evaluation framework MT-RAIG EVAL** that enables finer distinctions in assessing the quality of the long-form outputs. We focus on evaluating the following two key dimensions:

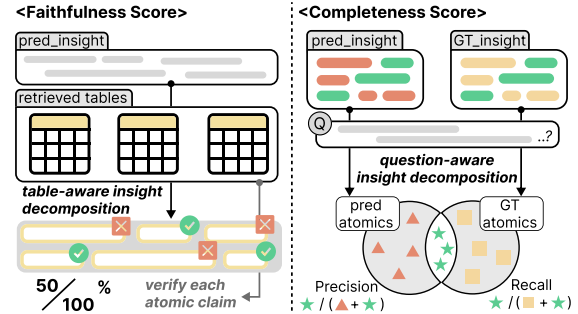


Figure 3: Overview of MT-RAIG EVAL

Faithfulness Score Faithfulness score evaluates whether an insight is fully grounded in the provided source tables. Ideally, a perfect evaluator should assess whether all atomic facts entangled within an insight are correctly grounded in the retrieved tables. To systematically verify these atomic facts, we introduce *table-aware insight decomposition*, which breaks down an insight into verifiable claims. Specifically, we leverage an LLM-based decomposer, enhanced with structural guidance from table schemas, to generate a set of granular claims explicitly linked to their originating tables. Each claim is then validated by an LLM verifier against the retrieved tables. This table-aware decomposition enables a fine-grained evaluation of the predicted insights by ensuring traceability and reducing ambiguity of the verifiable claims in multi-table contexts. Formally, given a set of retrieved tables \hat{T}_q , a predicted insight i , a decomposer \mathcal{D} , and a verifier \mathcal{V} , the final score $S_{Faith.}$ is computed as follows:

$$C = \mathcal{D}(\hat{T}_q, i), S_{Faith.} = \frac{1}{|C|} \sum_{k=1}^{|C|} \mathcal{V}(c_k, \hat{T}_q) \quad (2)$$

where $C = \{c_k\}_{k=1}^n$ is a set of decomposed claims, and $\mathcal{V}(c_k, \hat{T}_q) \in \{0, 1\}$ verifies each claim against the retrieved tables \hat{T}_q .

Completeness Score Completeness refers to the idea that a generated insight should fully address all the requirements outlined in the given question. Independent of faithfulness, this dimension penalizes outputs that omit key analytical steps or introduce redundant content that deviates from the

question’s intent. To verify that these key steps are properly addressed, we propose a *question-aware insight decomposition*, which deconstructs both the ground truth insight i and the predicted insight \hat{i} into atomic topics representing key steps necessary to resolve the question (*e.g.*, identifying causal relationships or synthesizing cross-table comparisons). Specifically, we first employ an LLM-based decomposer \mathcal{D} , conditioned on the input question q , to generate two sets of atomic topics $A = \mathcal{D}(q, i)$ and $\hat{A} = \mathcal{D}(q, \hat{i})$. Subsequently, we perform a semantic matching by using an LLM \mathcal{M} , between the atomic topics in A and \hat{A} to compute the precision P and recall R based on the degree of overlap between these sets. The final completeness score S_{Comp} is the F1 score, formulated as follows:

$$P = \frac{|\mathcal{M}(A, \hat{A})|}{|\hat{A}|}, R = \frac{|\mathcal{M}(A, \hat{A})|}{|A|}, F1 = \frac{2 \cdot P \cdot R}{P + R} \quad (3)$$

4 Experiments

4.1 Baselines

We benchmark the performance of diverse baselines on MT-RAIG BENCH. For multi-table retrieval, we consider both *general-purpose* retriever (Robertson and Zaragoza, 2009; Karpukhin et al., 2020; Izacard et al., 2021), and *table-specific* retriever (Herzig et al., 2021; Zhang et al., 2023) as the baselines. For insight generation, we employ 11 LLM-based baselines including *proprietary* (OpenAI, 2025b, 2024; Anthropic), *open-source* (DeepSeek-AI, 2025; Yang et al., 2024; Mesnard et al., 2024; MetaAI, 2024; Jiang et al., 2023), and *SOTA-TQA* methods (Ye et al., 2023; Wang et al., 2024b; Zhao et al., 2024) for comprehensive evaluations. The results are shown in Table 4 and Table 6, respectively. Please refer to Appendix B for the detailed information on the experimental setup and all the baselines.

4.2 Results

Multi-table Retrieval Evaluation To understand the challenges of multi-table retrieval in MT-RAIG BENCH, we compare the performance of five widely-used retrievers. From the results in Table 4, we observe that there remains a significant gap between high and low top- k performance, which is critical for the generation performance given that LLMs often hold only a limited number of tables in their context window. We also find that table-specific embedding models do not improve retrieval performance on MT-RAIG

BENCH. Instead, general text-based embedding model achieves better performance, similar to the findings of Wang et al. (2022). We attribute this to the nature of the MT-RAIG task, which prioritizes identifying insight-level semantic connection of tables for multi-hop queries rather than understanding the structural characteristics of tables. Based on these findings, we use DPR’s top-10 retrieved tables (considering the LLM’s context window limitation) for experiments in Table 5 and Table 6.

Type	Retriever	R@2	R@5	R@10	R@20
<i>general-purpose</i>	BM25	17.26	27.19	33.72	41.16
	DPR	44.58	68.24	80.83	88.45
	Contriever	23.47	35.67	44.27	52.12
<i>table-specific</i>	DTR	<u>37.77</u>	<u>59.60</u>	<u>74.50</u>	<u>86.22</u>
	TableLlama	36.93	59.48	72.44	81.56

Table 4: Multi-table retrieval results on MT-RAIG BENCH. We report Recall at various top- k . Full results including Precision and F1 are in Appendix D.

Meta Evaluation We conduct a meta evaluation to assess the reliability of MT-RAIG EVAL against existing automatic metrics. Specifically, we construct a meta evaluation dataset comprising 250 pairs of responses sampled from baseline generators, where each pair is labeled by two human evaluators based on their relative preferences on response faithfulness and completeness. For scoring, we follow the setting of Ru et al. (2024) to normalize each auto-evaluation score difference to the human preference scale of $[-1, 0, 1]$ and then measure the Pearson correlation with human preference ratings. Additionally, we report the correlation between the human evaluators as the upper bound. From the results in Table 5, we can observe that MT-RAIG EVAL achieves the highest correlation with human preference ratings across both dimensions, demonstrating its reliability over baseline methods for evaluating MT-RAIG BENCH. We provide more detailed results in Appendix D.

Evaluation Metric	Faith.	Comp.
SacreBLEU (Post, 2018)	31.33	33.01
ROUGE-L (Lin and Hovy, 2003)	27.69	43.43
METEOR (Banerjee and Lavie, 2005)	37.00	39.96
BERTScore (Zhang et al., 2019)	24.82	43.29
A3CU (Liu et al., 2023b)	41.22	40.10
TAPAS-Acc (Liu et al., 2022)	-10.40	21.29
G-Eval (Liu et al., 2023a)	<u>47.82</u>	26.35
MT-RAIG EVAL (ours)	64.94	67.67
Inter-Human Correlation	84.81	75.70

Table 5: Meta Evaluation results.

Generator	Faithfulness						Completeness					
	A&S	C&R	P&O	T&P	Avg. \hat{T}_q	Avg. T_q	A&S	C&R	P&O	T&P	Avg. \hat{T}_q	Avg. T_q
<i>Proprietary LLMs</i>												
o3-mini	38.62	42.11	35.12	39.57	38.85	42.57	62.84	58.78	59.77	59.42	60.20	60.81
GPT-4o	36.60	39.79	33.96	37.57	36.98	41.46	64.46	59.37	61.49	60.77	61.52	63.28
Claude 3.5 Sonnet	41.00	42.45	35.60	39.66	39.68	43.35	61.67	56.06	58.94	57.86	58.63	59.70
<i>Open-source LLMs</i>												
DeepSeek-R1-8B	37.05	36.57	34.09	34.51	35.55	40.12	64.23	58.72	61.41	59.48	60.96	63.41
Qwen2-7B	33.79	36.97	29.85	33.01	33.40	40.59	61.92	57.88	59.73	57.32	59.21	61.57
Gemma-7B	26.51	28.91	23.63	22.60	25.41	30.31	62.72	57.73	60.52	59.68	60.16	61.64
Llama 3.1-8B	32.30	33.71	29.73	30.63	31.59	37.62	59.20	55.11	57.45	55.85	56.90	58.14
Mistral-7B	31.33	33.11	27.58	29.81	31.33	36.26	62.96	58.00	59.66	58.29	59.73	62.77
<i>SOTA TQA-methods</i>												
Dater	26.33	30.26	27.06	28.04	27.92	32.43	59.70	57.26	60.35	59.16	59.12	62.32
Chain-of-Table	32.86	33.28	30.11	31.33	31.90	37.71	59.29	55.49	58.58	57.17	57.63	62.15
TaPERA	18.38	21.06	18.16	19.38	19.25	20.68	59.23	53.56	58.05	55.53	56.59	55.28

Table 6: A&S, C&R, P&O, and T&P correspond to the types of question in Table 2. Avg. \hat{T}_q denotes the average open-domain performance of different question types with retrieved tables, and Avg. T_q indicates closed-domain performance using only the ground truth tables. Precision/Recall for Completeness are detailed in Appendix D.

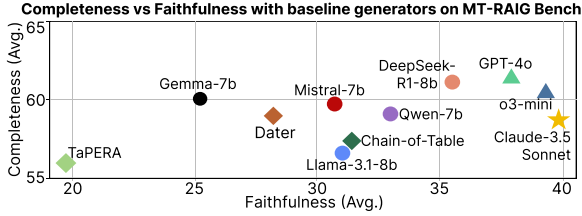


Figure 4: Comparison of various baseline generators based on their average \hat{T}_q completeness and faithfulness.

Insight Generation Evaluation Leveraging the MT-RAIG EVAL, we evaluate the insight generation performance of various baselines on MT-RAIG BENCH. From the results in Table 6 and Figure 4, we derive the following key conclusions: (1) **MT-RAIG BENCH poses a significant challenge for insight-level table reasoning, even for frontier LLMs.** Both open-source and proprietary models struggle to generate insights for questions in MT-RAIG BENCH, achieving only around 40% in faithfulness and 60% in completeness even when provided with gold tables T_q as input. (2) **Deep thinking with scaling test-time compute also shows promise in insight-level table reasoning tasks.** Although proprietary models generally outperform open-source counterparts, DeepSeek-R1-8b achieves performance on par with proprietary models and even surpasses Claude and o3-mini in completeness score, despite having a smaller number of parameters. Given that a similar reasoning model, o3-mini, also demonstrates strong performance, these results supports OpenAI (2025a); DeepSeek-AI (2025)’s recent finding that inference-time scaling substantially enhances models’ capac-

ity for complex reasoning, suggesting its applicability can also extend to insight-level table reasoning tasks. (3) **SOTA TQA-methods underperform compared to general-purpose LLMs.** We attribute this to the focus of existing TQA methods on single-table factoid extraction and their specialization in closed-domain scenarios, which limits their robustness to noise from irrelevant tables.

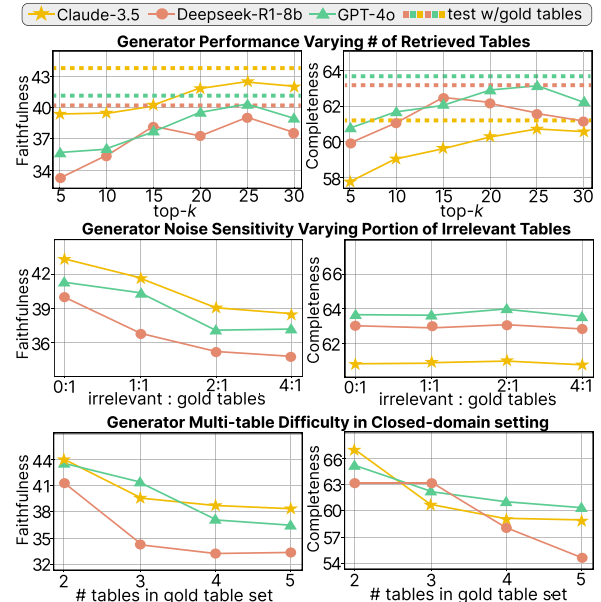


Figure 5: (Upper) Generator performance varying the number of retrieved tables. Dash lines are closed-domain performance serving as the upper bound for each model. (Middle) Generator noise sensitivity varying portion of irrelevant tables against gold tables. (Lower) Generator multi-table difficulty comparison according to the number of tables in the gold table set.

4.3 Further Analysis

Trade-offs in retrieving more tables We first analyze the relationship between retrieval and generation performance by varying the number of retrieved tables (k). While increasing the number of k is an intuitive way to improve generation performance in RAG systems (Kim et al., 2024a), we investigate whether this also holds for the table domain, where retrieved information is structurally more complex than general text. From the results in Figure 5 (Upper), we first observe that the generation performance broadly aligns with conventional intuition, exhibiting an overall improvement across all models as k increases. However, we also find that a continued increase in k does not guarantee sustained performance gains. Beyond a certain threshold, performance plateaus or even declines, suggesting that retrieving an excessive number of tables introduces noise into the generation process.

Factuality of the generated insights are more sensitive to the noisy tables Guided by the observations in Figure 5 (Upper), we further investigate how sensitively the generator responds to noisy table information. To isolate the impact of noise from retrieval errors, we analyze samples where the entire gold table set is successfully retrieved. Within these samples, we analyze the performance by varying the ratio of irrelevant to gold tables while keeping the gold tables fixed as part of the model input. The results in Figure 5 (Middle) reveal that, as the proportion of irrelevant tables increases, faithfulness shows a marked decrease while completeness remains relatively stable. This suggests that although including the correct tables ensures the generator can somewhat address the high-level key steps (i.e., maintain completeness), the added noise from irrelevant tables hinders the extraction of precise factors for the final insights.

Even without the noisy tables, the model still struggles as the number of tables to reference increases While our findings indicate that generation performance degrades with noisy tables, one might wonder how challenging it becomes for the model as the number of source tables it should handle increases, even in noise-free conditions. To explore this, we conduct a closed-domain experiment in which only the gold tables are provided as input, and investigate how generation performance varies depending on the number of tables in the gold table set that the model should reference. From the re-

sults in Figure 5 (Lower), we observe a significant drop in both faithfulness and completeness as the number of gold tables to reference grows, showing that current models struggle to conduct complex reasoning across multiple sources of tables.

5 Related Work

Benchmarks for Table-based Tasks Benchmarks for table-based reasoning has been studied across several major tasks, including text-to-SQL (Zhong et al., 2018), table-to-text (Parikh et al., 2020), and tableQA (Pasupat and Liang, 2015a). Early works (Yu et al., 2018; Nan et al., 2021) mainly focus on measuring the performance of neural models as executors for extracting factoid-level information, where models generate SQL queries or direct answers from tables. More recently, some works (Zhao et al., 2023a; Seo et al., 2024) target insight-level table reasoning tasks, moving beyond extracting explicit information from tables. Unlike prior works, our MT-RAIG BENCH significantly differs by exploring more realistic scenario where the system retrieves multiple evidence tables to generate insightful responses.

Automated Evaluation of Long-form Outputs in Table-based Tasks Existing methods have primarily compute the lexical overlap (Post, 2018), or semantic similarity with the reference (Zhang et al., 2019), while some (Liu et al., 2022) focus on factuality assessment using a trained verifier to check the grounding against the source table. Some recent studies (Zhao et al., 2024; Wang et al., 2024a) find that LLM-based evaluation metrics (i.e., G-Eval), show stronger alignment with human evaluation. Despite these advancements, these metrics still struggle to capture the fine-grained quality of long-form outputs. To address this gap, we propose a novel decomposition-based evaluation framework, MT-RAIG EVAL, which outperforms existing metrics in aligning with human judgments.

6 Conclusion

This work introduces MT-RAIG BENCH, the first large-scale benchmark for retrieval-augmented insight generation over multiple tables, alongside MT-RAIG EVAL, a novel automated evaluation framework designed to address the limitations of conventional metrics in assessing multi-table insights. Extensive experiments reveal that even frontier LLMs and SOTA TableQA systems struggle to meet the challenges posed by MT-RAIG BENCH.

Limitations

Although we believe that our MT-RAIG BENCH and MT-RAIG EVAL could serve as valuable resources for the research community, several limitations remain, suggesting areas for future improvement. First, while the synthetic generation of MT-RAIG BENCH enables scalable and consistent dataset construction (Tan et al., 2024; Hao et al., 2024; Han et al., 2024) as we discussed in Section 2, this approach carries inherent risks (Bender et al., 2021), such as reduced linguistic diversity in questions or potential overalignment with model generated responses that are unnatural or unfaithful description. Despite this downside, we minimize these risks by adopting a human-in-the-loop annotation process, where human experts iteratively refine machine-generated questions. Further, we conduct a machine-human dual quality control, in which automated filters first strictly discard low-quality instances, followed by human validation to ensure the quality of the final benchmark.

Second, although MT-RAIG BENCH covers both relational DB tables and Wikipedia tables that constitute the two main portion of table domain, it could be beneficial to add more data sources from diverse domains (e.g., financial, scientific, or medical tables). Future works could examine how existing models perform in these specialized domains to assess their robustness to domain-specific terminology or extremely large table sets.

Lastly, a notable limitation lies in our MT-RAIG EVAL’s reliance on LLM for assessment. While recent works increasingly adopt LLM-based evaluators in diverse domains due to their scalability and flexibility (Ye et al., 2024; Kim et al., 2024d; Ru et al., 2024; Han et al., 2024; Wang et al., 2024a), it is important to note that their judgments may still be affected by biases or inconsistencies inherent in the backbone model’s training data or architectural design. However, we believe this issue has been largely minimized in our study. Through a meta-evaluation in Section 3, we demonstrate our MT-RAIG EVAL’s reliability in aligning with human judgments. Furthermore, additional experiments that replacing the MT-RAIG EVAL’s LLM backbone with open-source models (Table 17) confirm the reproducibility of our evaluation framework.

Ethical Consideration

The output of text generation from LLMs may sometimes contain harmful, biased, or offensive

content and contain the risk of potential hallucination (Kim et al., 2024b). However, in our research, we assert that this risk is largely minimized. The source tables and source natural language facts used in the construction of our MT-RAIG BENCH are derived from SPIDER (Yu et al., 2018) and Open-WikiTable (Kweon et al., 2023), both of which are publicly available datasets licensed under the CC BY-SA 4.0, and these datasets have been annotated by human experts. Additionally, we manually reviewed the generated texts and eliminate any toxic, offensive, or biased language to ensure the quality and fairness of the auto-generated contents. For human evaluation, two graduate students participated as annotators, receiving a pre-guide before taking part in evaluation and validation. We advise the annotators to complete no more than 20 unit tasks per day, and the entire annotation process spanned approximately 30 days.

Acknowledgement

This work was supported by the IITP grants funded by the Korea government (MSIT) (No. RS-2020-II201361; RS-2024-00457882, AI Research Hub Project), and the NRF grant funded by the Korea government (MSIT) (No. RS-2025-00560295).

References

- Anthropic. [Claude 3.5 sonnet model card addendum](#).
- Satanjeev Banerjee and Alon Lavie. 2005. [Meteor: An automatic metric for mt evaluation with improved correlation with human judgments](#). In *IEEE Evaluation@ACL*.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Wenhu Chen. 2023. [Large language models are few\(1\)-shot table reasoners](#). *Preprint*, arXiv:2210.06710.
- Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, SHIYANG LI, Xiyu Zhou, and William Yang Wang. 2019. [Tabfact: A large-scale dataset for table-based fact verification](#). *ArXiv*, abs/1909.02164.
- Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. [HybridQA: A dataset of multi-hop question answering over tabular and textual data](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*,

- pages 1026–1036, Online. Association for Computational Linguistics.
- Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. [Decontextualization: Making sentences stand-alone](#). *Transactions of the Association for Computational Linguistics*, 9:447–461.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jinyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. [RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA. Association for Computational Linguistics.
- Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. [Synthetic data in ai: Challenges, applications, and ethical implications](#). *Preprint*, arXiv:2401.01629.
- Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. [Open domain question answering over tables via dense retrieval](#). In *North American Chapter of the Association for Computational Linguistics*.
- Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. [Tapas: Weakly supervised table parsing via pre-training](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. [Unsupervised dense information retrieval with contrastive learning](#). *Trans. Mach. Learn. Res.*, 2022.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L’elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). *ArXiv*, abs/2004.04906.
- Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024a. [Sure: Summarizing retrievals using answer candidates for open-domain QA of LLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Seoyeon Kim, Kwangwook Seo, Hyungjoo Chae, Jinyoung Yeo, and Dongha Lee. 2024b. [VerifiNER: Verification-augmented NER via knowledge-grounded reasoning with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2441–2461, Bangkok, Thailand. Association for Computational Linguistics.
- Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashtevski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2024c. [Evaluating language models as synthetic data generators](#). *Preprint*, arXiv:2412.03679.
- Sunghwan Kim, Tongyoung Kim, Kwangwook Seo, Jinyoung Yeo, and Dongha Lee. 2024d. [Stop playing the guessing game! target-free user simulation for evaluating conversational recommender systems](#). *Preprint*, arXiv:2411.16160.
- Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. 2023. [Open-wikitable: Dataset for open domain question answering with complex reasoning over table](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). *Preprint*, arXiv:2309.06180.
- Chin-Yew Lin and Eduard H. Hovy. 2003. [Automatic evaluation of summaries using n-gram co-occurrence statistics](#). In *North American Chapter of the Association for Computational Linguistics*.
- Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. 2022. [Plog: Table-to-logic pre-training for logical table-to-text generation](#). *ArXiv*, abs/2205.12697.
- Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023a. [G-eval: Nlg evaluation using gpt-4 with better human alignment](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. [Towards interpretable and efficient automatic reference-based summarization evaluation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.
- Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Rivière, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L’eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone,

- Am'elie H'eliou, Andrea Tacchetti, and 88 others. 2024. [Gemma: Open models based on gemini research and technology](#). *ArXiv*, abs/2403.08295.
- MetaAI. 2024. [Introducing Meta Llama 3: The most capable openly available llm to date](#).
- Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. [Scigen: a dataset for reasoning-aware text generation from scientific tables](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Linyong Nan, Chia-Hsuan Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Benjamin Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2021. [Fetaqa: Free-form table question answering](#). *Transactions of the Association for Computational Linguistics*, 10:35–49.
- Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. [Mixeval: Deriving wisdom of the crowd from LLM benchmark mixtures](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- OpenAI. 2024. [Gpt-4o system card](#). *arXiv*:2410.21276.
- OpenAI. 2025a. [learning-to-reason-with-llms](#).
- OpenAI. 2025b. [Openai o3-mini system card](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. [Totto: A controlled table-to-text generation dataset](#). *ArXiv*, abs/2004.14373.
- Panupong Pasupat and Percy Liang. 2015a. [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 (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.
- Panupong Pasupat and Percy Liang. 2015b. [Compositional semantic parsing on semi-structured tables](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Matt Post. 2018. [A call for clarity in reporting bleu scores](#). In *Conference on Machine Translation*.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. [In-context retrieval-augmented language models](#). *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- Stephen E. Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3:333–389.
- Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. [RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Kwangwook Seo, Jinyoung Yeo, and Dongha Lee. 2024. [Unveiling implicit table knowledge with question-then-pinpoint reasoner for insightful table summarization](#). *ArXiv*, abs/2406.12269.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. [REPLUG: Retrieval-augmented black-box language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for data annotation and synthesis: A survey](#). *Preprint*, *arXiv*:2402.13446.
- Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gestein. 2024. [Struc-bench: Are large language models really good at generating complex structured data?](#) *Preprint*, *arXiv*:2309.08963.
- Yixuan Tang and Yi Yang. 2024. [Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries](#). In *First Conference on Language Modeling*.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *ArXiv*, abs/2307.09288.
- Yuqi Wang, Lyuhao Chen, Songcheng Cai, Zhijian Xu, and Yilun Zhao. 2024a. [Revisiting automated evaluation for long-form table question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14696–14706, Miami, Florida, USA. Association for Computational Linguistics.
- Zhiruo Wang, Zhengbao Jiang, Eric Nyberg, and Graham Neubig. 2022. [Table retrieval may not necessitate table-specific model design](#). In *Proceedings of*

- the Workshop on Structured and Unstructured Knowledge Integration (SUKI)*, pages 36–46, Seattle, USA. Association for Computational Linguistics.
- Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024b. [Chain-of-table: Evolving tables in the reasoning chain for table understanding](#). *ArXiv*, abs/2401.04398.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. [Long-form factuality in large language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 39 others. 2024. [Qwen2 technical report](#). *ArXiv*, abs/2407.10671.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. [\$\tau\$ -bench: A benchmark for tool-agent-user interaction in real-world domains](#). *Preprint*, arXiv:2406.12045.
- Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. [FLASK: Fine-grained language model evaluation based on alignment skill sets](#). In *The Twelfth International Conference on Learning Representations*.
- Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. [Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning](#). *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Tao Yu, Rui Zhang, Kai-Chou Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Z Li, Qingning Yao, Shanell Roman, Zilin Zhang, and Dragomir R. Radev. 2018. [Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task](#). *ArXiv*, abs/1809.08887.
- Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2023. [Tablellama: Towards open large generalist models for tables](#). In *North American Chapter of the Association for Computational Linguistics*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *ArXiv*, abs/1904.09675.
- Yilun Zhao, Lyuhao Chen, Arman Cohan, and Chen Zhao. 2024. [TaPERA: Enhancing faithfulness and interpretability in long-form table QA by content planning and execution-based reasoning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12824–12840. Association for Computational Linguistics.
- Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Xiangru Tang, Yumo Xu, Arman Cohan, and Dragomir R. Radev. 2023a. [Qtsumm: A new benchmark for query-focused table summarization](#). *ArXiv*, abs/2305.14303.
- Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023b. [Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.
- Victor Zhong, Caiming Xiong, and Richard Socher. 2018. [Seq2SQL: Generating structured queries from natural language using reinforcement learning](#).

A Benchmark Construction Details

A.1 Data Annotation

We show prompts used in data annotation process from Table 10 to 13.

Multi-table Set Collection To extend the existing single table corpus into a collection of multi-table sets, we classify scenarios where multiple tables are combined into two categories:

- *Joinable tables*: We first classify the SQL queries in the existing dataset SPIDER based on the join operator to identify tables that can be joined and then link the tables as the multi-table set.
- *Topic-related tables*: To cluster topically related tables, we semantically group the tables with the table meta data. Within each grouped table set, all tables are linked with related topics (at least two matching subtitles among page, section, or caption titles) and similar headers (differing by at most one column name). To consider the spatio-temporal relationships between tables, we use exact matching while excluding numeric values.

Question Types We provide the following definitions for each type of question to human reviewers who guide the agent with initial seed questions and validate the generated samples iteratively:

- *Analysis & Summary*: Synthesize multiple data sources into a coherent narrative, focusing on interpretation rather than raw figures. Identifies key metrics and contextualizes quantitative outcomes to uncover biases influencing decision-making.
- *Performance & Outcome*: Evaluate measurable achievements by linking performance metrics to qualitative success factors. Assesses growth and long-term accomplishments to provide a broader context for understanding progress.
- *Comparison & Relationship*: Analyze relationships between data points, explaining connections between attributes. Examines structural components and grouping logic to reveal organizational hierarchies and patterns, enhancing understanding of data organization.
- *Trend & Pattern*: Examine temporal or categorical changes to identify recurring behaviors and systemic shifts. Goes beyond documenting events to interpret transformations, providing insights into structural changes or cyclical phenomena.

A.2 Quality Control

Agent Self-Verification We first utilize an agent annotator as a self-verifier, ensuring that each multi-table set, question, and insight triple satisfies strict criteria:

- **Relevance**: Does the question appropriately capture the relationships between the tables in the multi-table set, ensuring that it pertains to all tables and can be answered solely using the provided information?
- **Faithfulness**: Does the insight accurately reflect the information within the multi-table set, ensuring it remains grounded in the given data while providing a clear and unambiguous response?
- **Completeness**: Does the insight fully and logically address the question, covering all necessary aspects while maintaining clarity and coherence?

Any triple (multi table set, question, and insight) that fails at least one of these standards is discarded. Representative examples of discarded data are in Table 24 and 25. We report the statistics of self-verification process in Table 7. Prompts used in agent verification process for three criteria are shown in Table 14.

Ratio	Relevance	Faithfulness	Completeness
%	64.73	19.38	7.47

Table 7: Discarded data ratio for each criterion in self-verification process. Only triples that pass all these standards are used as MT-RAIG BENCH.

Human Validation Alongside the agent verification, we incorporate human validation as critical review points for the annotated data, balancing the efficiency of automated annotation with the reliability of human annotation to ensure the quality of MT-RAIG BENCH. We provide the human validation interface in Figure 6. Specifically, we adopt the following criteria to check the quality of table, question, and insight triples:

- **Inter-Table Relevance**: Measures how effectively data from different tables can be connected or combined to provide comprehensive insights.
- **Table-Question Relevance**: Assesses the extent to which the contents of a table directly support and address the given question.
- **Question Complexity**: Evaluates the level of difficulty and the number of factors or layers involved

in understanding or answering the question.

- **Question Meaningfulness**: Determines whether the question is clearly defined, significant, and natural.
- **Question-Insight Completeness**: Checks if the insight provides all the necessary key information to address the question.
- **Table-Insight Faithfulness**: Ensures that the insights drawn accurately and reliably reflect the source tables.

B Experimental Details

B.1 Meta Evaluation

Considering that all automatic metrics are designed with distinct objectives and functionalities, direct numerical comparisons between their scores are inherently limited. Instead, it is intuitive that a metric’s reliability should be judged by its capacity to align with relative human preferences. To operationalize this principle, we adopt the meta-evaluation protocol proposed by [Ru et al. \(2024\)](#), constructing a meta evaluation dataset of 250 response pairs sampled from 11 different baseline generators. Human evaluators are then tasked with annotating each pair across two dimensions—faithfulness and completeness—by selecting one of three options: win, tie, or loss. The annotation interface is illustrated in Figure 7. We implement MT-RAIG EVAL on the top of `gpt-4o-mini-2024-07-18`. To check the reproducibility of MT-RAIG EVAL, we additionally conduct an ablation study in Table 21, examining the impact of different backbone LLMs on MT-RAIG EVAL. We consider the following automatic metrics as baselines:

- **SacreBLEU** ([Post, 2018](#)) standardizes BLEU ([Papineni et al., 2002](#)) score calculations by ensuring consistent and reproducible results. It measures the geometric mean of n-gram precision over the output text.
- **ROUGE-L** ([Lin and Hovy, 2003](#)) evaluates text similarity based on the longest common subsequence. Specifically, we reported F1 score.
- **METEOR** ([Banerjee and Lavie, 2005](#)) evaluates text similarity by using unigram matching between machine-generated outputs and human reference texts.
- **BERTScore** ([Zhang et al., 2019](#)) computes the similarity between generated and reference texts

using contextual word embeddings.

- **A3CU** ([Liu et al., 2023b](#)) evaluates summarization quality by directly comparing texts without extracting atomic content units, providing a human-aligned assessment of content similarity.
- **TAPAS-Acc** ([Liu et al., 2022](#)) is a reference-free metric that uses a TAPAS ([Herzig et al., 2020](#)) model fine-tuned on the TabFact ([Chen et al., 2019](#)) dataset to assess the faithfulness of generated text by verifying factual consistency.
- **G-Eval** ([Liu et al., 2023a](#)) assesses the quality of generated text based on specific evaluation criteria using LLMs. We adopt G-Eval to separately evaluate *faithfulness* and *completeness* on a 5-point Likert scale. The evaluation prompts are provided in Table 18.

B.2 Retrieval Baselines

To assess the effectiveness of multi-table retrieval, we evaluate both general-purpose retrievers and table-specific retrievers. General-purpose retrievers serve as strong baselines, given their widespread use in plain retrieval tasks. Table-specific retrievers are designed explicitly for tables, leveraging table-specific representations for improved retrieval performance. Specifically, we consider the following retrievers as the baselines:

General-purpose retrievers

- **BM25** ([Robertson and Zaragoza, 2009](#)) is a sparse retriever that relies on a traditional bag-of-words representation to score relevant documents. In this study, we rank linearized tables as documents based on term frequency and inverse document frequency, leveraging lexical overlaps between questions and tables.
- **DPR** ([Karpukhin et al., 2020](#)) employs BERT-based encoders to independently map questions and documents into a shared embedding space. It learns dense vector representations that enable semantic similarity matching. In this study, we implement DPR utilizing `facebook/dpr-encoder-multiset-base`.
- **Contriever** ([Izacard et al., 2021](#)) encodes questions and documents into a shared embedding space, optimizing for relevance through contrastive learning. In this study, we implement Contriever utilizing `facebook/contriever-msmacro`.

Table-specific retrievers

- **DTR** (Herzig et al., 2021) is a table-specific dense retriever built on a TAPAS (Herzig et al., 2020) backbone, designed to effectively encode tabular structures and relationships. In this study, we fine-tune google/tapas-base.
- **TableLlama** (Zhang et al., 2023) is Llama 2-7B (Touvron et al., 2023) based open-source LLM-based generalist model that is designed for a variety of table-based tasks. We utilize TableLlama to generate question and table embeddings for table retrieval task. In this study, we fine-tune osunlp/TableLlama.

B.3 Generator Baselines

To comprehensively compare the performance of insight generation across multiple tables, we consider three types of baselines: proprietary LLMs, open-source LLMs, and SOTA table question-answering methods. Proprietary LLMs encompass commercial models renowned for their advanced reasoning and high-quality text generation capabilities. Open-source LLMs, serving as transparent and adaptable alternatives, leverage large-scale training data and advanced architectures to deliver performance that is competitive with proprietary counterparts. SOTA TQA methods are specialized approaches explicitly engineered for complex, table-aware reasoning tasks, prioritizing accuracy in tabular question-answering. We consider the following models as the baseline generators:

Proprietary LLMs

- **o3-mini** (OpenAI, 2025b) is a smaller yet advanced LLM developed by OpenAI, which is designed to efficiently solve complex problems by breaking them into constituent parts. In this study, we leverage o3-mini-2025-01-31 with reasoning effort parameter as medium.
- **GPT-4o** (OpenAI, 2024) is an advanced OpenAI’s proprietary LLM known for its enhanced reasoning capabilities and performance across various disciplines. In this study, we leverage gpt-4o-2024-08-06 checkpoint.
- **Claude 3.5 Sonnet** (Anthropic) is developed by Anthropic. It features improvements in coding proficiency and multimodal capabilities. In this study, we leverage claude-3-5-sonnet-20241022 checkpoint.

Open-source LLMs

- **DeepSeek-R1-8B** (DeepSeek-AI, 2025) is dis-

tilled version of DeepSeek-R1, which is an open-source LLM released by DeepSeek AI. In this study, we utilize unsloth/DeepSeek-R1-Distill-Llama-8B., which is based on Llama-3.1-8B.

- **Qwen2-7B** (Yang et al., 2024) is an open-source LLM developed by Alibaba Cloud, with the largest model containing 72 billion parameters. In this study, we utilize Qwen/Qwen2-7B-Instruct.
- **Gemma-7B** (Mesnard et al., 2024) is an open-source LLM developed by Google DeepMind, which is known for its multilingual capabilities and creative outputs. In this study, we utilize google/gemma-7b-it.
- **Llama 3.1-8B** (MetaAI, 2024) is an open-source LLM released by Meta AI, which is available in multiple sizes up to 405 billion parameters. In this study, we utilize meta-llama/Llama-3.1-8B-Instruct.
- **Mistral-7B** (Jiang et al., 2023) is an open-source LLM developed by Mistral AI. In this study, we utilize mistralai/Mistral-7B-Instruct-v0.3.

SOTA TQA-methods

- **Chain-of-Table** (Wang et al., 2024b) is LLM-based method designed to enhance table-based reasoning. It employs iterative reconstruction of input table through dynamic tabular operations.
- **TaPERA** (Zhao et al., 2024) a modular framework designed to enhance faithfulness and interpretability in long-form table question answering by combining a QA-based content planner and execution-based reasoning..
- **Dater** (Ye et al., 2023) focuses on selecting relevant information from input tables and providing contextual information to support the statement verification process.

B.4 Implementation Details

Table Input Serialization Following recent studies that utilize language models for table-related tasks (Chen, 2023; Seo et al., 2024), we serialize the table input into a flattened sequence to effectively represent table data for language model processing. The table title is enclosed within [TITLE] tags, followed by the table headers marked with [HEADER], where individual column names are separated by a vertical bar (|). Each row

is prefixed with a [ROW] tag and an index, while cell values are separated by a vertical bar. This approach ensures that the table format is preserved while making the input compatible with language models. For example, the input table is formatted as follows:

```
[TITLE] title [HEADER] col 1 |
col2 | ... [ROW 1] cell 1,1 | cell
1,2 | ... [ROW 2] cell 2,1 | ...
```

Training and Inference For the table-specific retriever, we fine-tune each model on a single-table QA dataset using the AdamW optimizer with a learning rate of for 5 epochs. To efficiently fine-tune the model, we adopt LoRA and set the parameters as $r = 8$, $\alpha = 32$. We use a constant learning rate schedule set at $2e-5$, and train with the batch size of 1 on 4 NVIDIA A100 GPU. The inference of open-source LLMs are conducted using vLLM framework (Kwon et al., 2023). We set temperature to 0.0 for efficient and robust output generation.

C Ablation Study

C.1 Reproducibility of MT-RAIG EVAL Backbone

To assess whether MT-RAIG EVAL produces consistent results across different backbone LLMs, we analyze two open-source models (DeepSeek-R1-8B and Llama-3.1-8B) against the GPT-4o-mini, which is the original backbone of MT-RAIG EVAL. Specifically, we measure how closely the evaluation scores from each open-source model align with those from GPT-4o-mini by calculating pairwise correlation. From the results in Table 8, we can observe that even when replacing the backbone with open-source models, the high correlation persists, demonstrating that our evaluation method is both reproducible and robust.

Backbone	Faithfulness	Completeness
DeepSeek R1-8B	74.12	77.85
Llama-3.1-8B	80.78	72.82

Table 8: Pearson correlation scores varying MT-RAIG EVAL backbone compared to GPT-4o mini.

C.2 Effect of Parameter Scaling on MT-RAIG BENCH Performance

From the results in Table 6, we observe that proprietary models with larger model sizes generally show higher performance compared to open-source

models that have relatively fewer parameters. To further investigate this, we conduct additional experiments by differ the parameter sizes of open-source LLM to understand the effect of model size in the table reasoning performance. Specifically, we evaluate DeepSeek-R1, which is most powerful open-source LLM among our baselines ranging from 8B to 70B parameters. From the results in Table 9, we confirm a general trend where increasing model size correlates with improved performance, with a particularly notable improvement in the faithfulness score. This discrepancy suggests that the model’s capacity to accurately interpret and reason about table structures (which is a prerequisite for faithfulness) is more directly enhanced by parameter scaling, compared to the broader task coverage implied by completeness.

DeepSeek-R1	Faithfulness		Completeness	
	Score	$\Delta\%$	Score	$\Delta\%$
Distill-Llama-8B	35.55	-	60.96	-
Distill-Qwen-14B	37.26	+ 4.81	61.07	+ 0.18
Distill-Qwen-32B	39.02	+ 9.76	60.62	- 0.56
Distill-Llama-70B	38.11	+ 7.20	62.13	+ 1.92

Table 9: MT-RAIG EVAL results varying parameter size of DeepSeek-R1. $\Delta\%$ denotes the relative improvement in performance compared to the Distill-Llama-8B.

D Detailed Experimental Results

To supplement Table 5, we present the Spearman correlation results alongside the Pearson correlation, in terms of faithfulness and completeness, as shown in Table 17. We report the full multi-table retrieval results in Table 19, MT-RAIG EVAL Completeness score results of diverse baselines in Table 20, with respect to Precision, Recall, and F1 score across different top- k values. Additionally, we present insight generation results along with existing automatic metrics in Table 21.

E Case Study

We select representative examples of (1) MT-RAIG BENCH in Table 22 and 23, (2) discarded data in Table 24 and 25, (3) MT-RAIG EVAL Faithfulness and (4) Completeness evaluation in Table 26 and 27, respectively. Moreover, we present (5) comparison with MT-RAIG EVAL results for each baseline in Table 28 and 29. Ensure that we display only the first three rows of each table to improve readability.

Question Annotation

[Instruction]

You are a helpful assistant that generate questions based on keywords derived from the table titles and overlapping values that contain information about relationships between tables. Please refer to the given examples to capture the insight-level pattern, and generate a total of 10 questions.

[Inputs]

Examples: {seed_questions}

Keywords: {key_words_derived_from_table_titles}

Overlapping values: {overlapping_values}

[Output]

Questions:

Table 10: Prompt used for question annotation process.

Multi-Table Facts Expansion

[Instruction]

You are a helpful assistant that write a Python code to extract enriched facts from the given natural language facts based on table schemas. Please write a Python code in the following format.

```
```python
import pandas as pd # can import additional python libraries if necessary

def extract_patterns(dataframe, source_facts):
 """CODE"""
 return patterns

def expand_facts(dataframe, patterns):
 """Expand facts

 [Param]
 dataframe : pd.DataFrame
 patterns : List[str]

 [Return]
 expanded_facts_list : List[str]
 """
 """CODE_WITH_ITERATION"""
 return expanded_facts_list
```
```

[Inputs]

Table schemas: {table_schemas}

NL facts: {natural_language_facts}

[Output]

Python code:

Table 11: Prompt used for multi-table facts expansion process.

Question-Relevant Knowledge Extraction

[Instruction]

You are a helpful assistant that extract knowledge from natural language context that are relevant to the given question. Please filter out irrelevant content and extract question-relevant knowledge.

[Inputs]

Question: {question}

NL context: {expanded_natural_language_facts}

[Output]

Question-relevant knowledge:

Table 12: Prompt used for question-relevant knowledge extraction process.

Insight Annotation

[Instruction]

You are a helpful assistant that generate an insight to answer the given question using knowledge. Please generate an insight.

[Inputs]

Question: {question}

Knowledge: {question_relevant_knowledge}

[Output]

Insight:

Table 13: Prompt used for insight annotation process.

Self-Verification : Relevance

[Instruction]

You are a strict verifier for evaluate relevance. Please verify whether the given multi-table set and question pair is relevant based on the given criteria.

Criterion 1: Does the question focus on the relationships between the tables in the multi-table set?

Criterion 2: Does the question relate to all the tables in the multi-table set?

Criterion 3: Does the question rely solely on the information in the multi-table set for an answer?

Criterion 4: Does the question avoid requiring information that is not present in the multi-table set?

[Inputs]

Multi-table set: {serialized_tables}

Question: {question}

[Output]

Verification:

Self-Verification : Faithfulness

[Instruction]

You are a strict verifier for evaluate faithfulness. Please verify whether the given multi-table set and insight pair is faithful based on the given criteria.

Criterion 1: Does the insight relate to all tables within the multi-table set?

Criterion 2: Does the insight rely solely on the information within the multi-table set?

Criterion 3: Does the insight include all necessary details from the multi-table set?

Criterion 4: Does the insight provide a clear and unambiguous response based on the multi-table set?

[Inputs]

Multi-table set: {serialized_tables}

Insight: {insight}

[Output]

Verification:

Self-Verification : Completeness

[Instruction]

You are a strict verifier for evaluate completeness. Please verify whether the given question and insight pair is complete based on the given criteria.

Criterion 1: Does the insight directly relate to the question?

Criterion 2: Does the insight fully address all aspects of the question?

Criterion 3: Does the insight unfold logically and without contradictions in response to the question?

Criterion 4: Does the insight remain clear and understandable in the context of the question?

[Inputs]

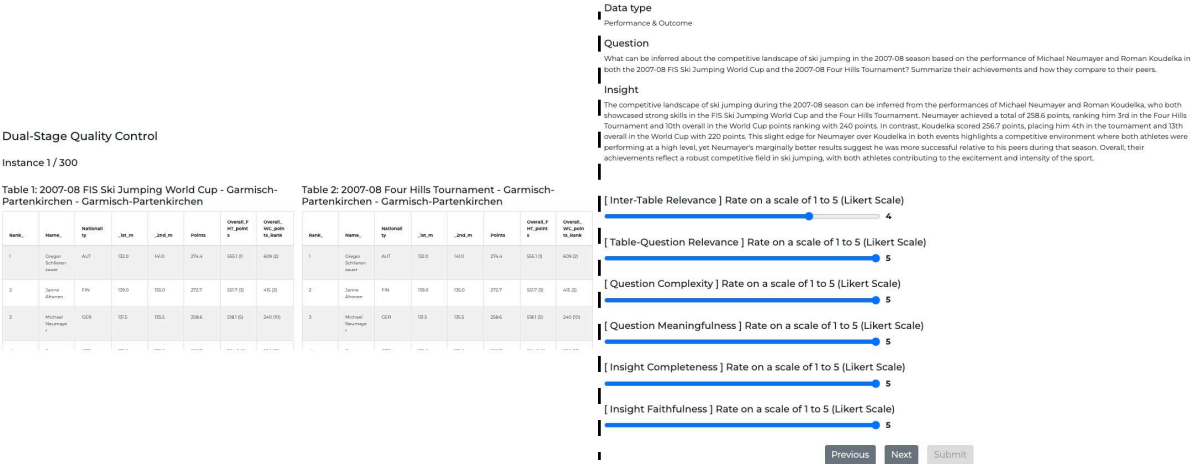
Question: {question}

Insight: {insight}

[Output]

Verification:

Table 14: Prompts used for agent verification process.



MT-RAIG Eval - Completeness : Question-aware insight decomposition

[Instruction]

You are a helpful assistant tasked with decomposing a given insight based on a question. Your goal is to extract atomic-level topics from the given insight.

[Inputs]

Question: {question}

Insight: {insight}

[Output]

Decomposed topic set:

MT-RAIG Eval - Completeness : Topic matching

[Instruction]

You will be given two topic sets. Your task is to match topics bidirectionally based on their semantic similarity, ensuring that both sets are treated with equal importance.

[Inputs]

Topic set A: {decomposed_ground_truth_topic_set}

Topic set B: {decomposed_predicted_topic_set}

[Outputs]

Matched topic subset of A:

Matched topic subset of B:

Table 16: Prompts for evaluate MT-RAIG EVAL Completeness.

Meta Evaluation Dataset

Instance 1 / 250

Question

Summarize the basic information of the herbaria located in major cities across the Baltic states, British Isles, and Spain and Portugal, focusing on their names and the number of specimens they hold.

Ground-truth insight

The herbaria located in major cities across the Baltic states, British Isles, and Spain and Portugal vary significantly in the number of specimens they hold. In the Baltic states, notable herbaria include the Estonian University of Life Sciences in Tartu with 336,000 specimens, the University of Tartu also in Tartu with 324,000 specimens, and Vilnius University in Vilnius with 300,000 specimens. Other significant collections are found at the University of Latvia in Riga with 176,000 specimens and the Estonian Museum of Natural History in Tallinn with 100,000 specimens. In the British Isles, the Royal Botanic Gardens Kew leads with an impressive 7,000,000 specimens, followed by the British Museum of Natural History in London with 3,200,000 specimens, and the Royal Botanic Garden in Edinburgh with 2,000,000 specimens. Spain and Portugal feature prominent herbaria such as the Real Jardín Botánico in Madrid with 850,000 specimens, the University of Coimbra in Coimbra with 800,000 specimens, and the Institut Botànic de Barcelona in Barcelona with 700,000 specimens. Other notable collections include the Instituto de Investigação Científica Tropical in Lisbon with 300,000 specimens and several universities in Spain holding between 50,000 and 300,000 specimens.

Response 1

The herbaria in major cities across the Baltic states, British Isles, and Spain and Portugal showcase a diverse range of specimen collections, with significant variations in size and distribution. In the Baltic states, the largest herbaria are found in Tartu, Estonia, with the Estonian University of Life Science and University of Tartu holding over 300,000 specimens each. Vilnius, Lithuania, also boasts a substantial collection at Vilnius University with 300,000 specimens. The British Isles house some of the world's most extensive herbaria, with the Royal Botanic Gardens Kew in London leading with an impressive 7 million specimens, followed by the British Museum of Natural History with 3.2 million. Edinburgh and Cambridge also have notable collections exceeding 1 million specimens each. In Spain and Portugal, the largest herbaria are located in capital cities, with Madrid's Real Jardín Botánico holding 850,000 specimens and Lisbon's Instituto de Investigação Científica Tropical and Museu Nacional de História Natural collectively housing 350,000 specimens. Barcelona also stands out with two significant herbaria totaling 1 million specimens. This distribution of herbaria highlights the importance of major urban centers as hubs for botanical research and preservation, with capital cities and renowned academic institutions often hosting the most extensive collections.

Response 2

The herbaria located in major cities across the Baltic states, British Isles, and Spain and Portugal are as follows: 1. In the Baltic states, the largest herbarium is the Estonian University of Life Sciences, which has 336,000 specimens. The second largest is the University of Tartu, with 324,000 specimens. The third largest is the Vilnius University, with 300,000 specimens. The fourth largest is the University of Latvia, Riga, with 176,000 specimens. The fifth largest is the University of Latvia, Salaspils, with 75,000 specimens. The sixth largest is the Institute of Botany, Lithuania, with 130,000 specimens. The seventh largest is the Estonian Museum of Natural History, with 100,000 specimens. The eighth largest is the Tallinn Botanic Garden, with 85,000 specimens. The ninth largest is the National Botanic Garden of Latvia, with 50,000 specimens. 2. In the British Isles, the largest herbarium is the Royal Botanic Gardens Kew, with 7,000,000 specimens. The second largest is the British Museum of Natural History, with 3,200,000 specimens. The third largest is the Royal Botanic Garden, Edinburgh, with 2,000,000 specimens. The fourth largest is the University of Cambridge, with 1,000,000 specimens. The fifth largest is the University of Manchester, with 1,000,000 specimens. The sixth largest is the National Botanic Gardens, Kew, with 850,000 specimens. The seventh largest is the National Museums & Galleries of Wales, with 550,000 specimens. The eighth largest is the Fielding-Druce Herbarium, University of Oxford, with 500,000 specimens. The ninth largest is the CABI Bioscience UK Centre, with 385,000 specimens. The tenth largest is the University of Museum Liverpool, with 350,000 specimens. The eleventh largest is the Daulby Herbarium, University of Oxford, with 300,000 specimens. The twelfth largest is the University of Reading, with 250,000 specimens. The thirteenth largest is Trinity College, with 230,000 specimens. The fourteenth largest is the University of Birmingham, with 200,000 specimens. The fifteenth largest is the University of Leicester, with 150,000 specimens. The sixteenth largest is the University of Aberdeen, with 120,000 specimens. The seventeenth largest is the South London Botanical Institute, with 100,000 specimens. The eighteenth largest is the Nottingham

In terms of faithfulness, does response1 win or loss compared to response2?

win (-1)

tie (0)

loss (1)

In terms of completeness, does response1 win or loss compared to response2?

win (-1)

tie (0)

loss (1)

Previous Next Submit

Retrieved Labels

Instance 1 / 250

Table 1: List of herbaria in Europe - Baltic states - Herbaria in the Baltic states

| NAME | No. Specimens | Addr. | Country | City |
|--------------------------------------|---------------|-------|-----------|-----------|
| Estonian University of Life Sciences | 336000 | TSL | Estonia | Tartu |
| University of Tartu | 324000 | TU | Estonia | Tartu |
| Vilnius University | 300000 | VI | Lithuania | Vilnius |
| University of Latvia - Riga | 176000 | RIG | Latvia | Riga |
| University of Latvia - Salaspils | 75000 | LATV | Latvia | Salaspils |
| Institute of Botany, Lithuania | 130000 | IBUG | Lithuania | Vilnius |

Table 2: List of herbaria in Europe - British Isles - Herbaria in the British Isles

| NAME | No. Specimens | Addr. | Country | City |
|-----------------------------------|---------------|-------|---------|---------------------|
| Royal Botanic Gardens Kew | 7000000 | RK | UK | Kew, England |
| British Museum of Natural History | 3200000 | BH | UK | London, England |
| Royal Botanic Gardens, Edinburgh | 2000000 | EB | UK | Edinburgh, Scotland |
| University of Cambridge | 1000000 | CC | UK | Cambridge, England |
| University of Manchester | 1000000 | MANCH | UK | Manchester, England |
| National Botanic Gardens, Ireland | 500000 | IBH | Ireland | Dublin |

Table 3: List of herbaria in Europe - Spain and Portugal - Herbaria in Spain and Portugal

| NAME | No. Specimens | Addr. | Country | City |
|---|---------------|-------|----------|-----------|
| Real Jardín Botánico | 850000 | RJB | Spain | Madrid |
| University of Coimbra | 800000 | CCO | Portugal | Coimbra |
| Institut Botànic de Barcelona | 700000 | IBC | Spain | Barcelona |
| Instituto de Investigação Científica Tropical | 300000 | IICCT | Portugal | Lisbon |
| Universitat de Barcelona | 300000 | UBN | Spain | Barcelona |
| Universitat de Sevilla | 200000 | USV | Spain | Seville |

Table 4: tracking_grants_for_research - organisations

| organisation_id | organisation_type | organisation_details |
|-----------------|-------------------|----------------------|
| 1 | RES | at |

Figure 7: Human annotation interface of the meta evaluation dataset.

| Metric | Faithfulness | | Completeness | |
|-------------------------|--------------|--------------|--------------|--------------|
| | Pearson | Spearman | Pearson | Spearman |
| SacreBLEU | 31.33 | 29.11 | 33.01 | 35.80 |
| ROUGE-L | 27.69 | 25.78 | <u>43.43</u> | 41.78 |
| METEOR | 37.00 | 36.54 | <u>39.96</u> | 37.57 |
| BERTScore | 24.82 | 37.43 | 43.29 | <u>42.10</u> |
| A3CU | 41.22 | 41.26 | 40.10 | 39.80 |
| TAPAS-Acc | -10.40 | -11.17 | 21.29 | 21.64 |
| G-Eval | <u>47.82</u> | <u>46.15</u> | 26.35 | 23.30 |
| MT-RAIG Eval | 64.94 | 65.47 | 67.67 | 68.97 |
| Inter-Human Correlation | 84.81 | 83.02 | 75.70 | 75.63 |

Table 17: Pearson and Spearman correlation results in terms of faithfulness and completeness.

G-Eval : Faithfulness

[Instruction]

You will be given retrieved tables, a question, and an insight. Your task is to evaluate how faithful the insight is to the data in the retrieved tables when answering the question. Please make sure you read and understand these instructions carefully.

Evaluation Criteria:

Faithfulness (1–5) - A good insight must be consistent with the retrieved tables and must not include contradictory or fabricated information.

Evaluation Steps:

1. Carefully review the retrieved tables and the question, noting all critical data.
2. Check if the insight strictly relies on the table data without contradicting any key details.
3. Assign a Faithfulness score (1–5) based on the insight’s accuracy and adherence to the tables.

[Inputs]

Retrieved tables: {serialized_tables}

Question: {question}

Insight: {insight}

[Output]

Faithfulness (scores ONLY):

G-Eval : Completeness

[Instruction]

You will be given retrieved tables, a question, and an insight. Your task is to evaluate whether the insight fully addresses the question using the data provided. Please make sure you read and understand these instructions carefully.

Evaluation Criteria:

Completeness (1–5) - A good insight must address all relevant elements of the question, leaving no key points from the tables unaddressed.

Evaluation Steps:

1. Review the retrieved tables and question to identify the essential information.
2. Compare the insight against the question and tables, checking if all crucial points are covered.
3. Assign a Completeness score (1–5) based on how thoroughly the insight uses the table data to fulfill the question.

[Inputs]

Retrieved tables: {serialized_tables}

Question: {question}

Insight: {insight}

[Output]

Completeness (scores ONLY):

Table 18: Prompts used for G-Eval.

| Retriever | Top-2 | | | Top-5 | | | Top-10 | | | Top-20 | | |
|-----------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| <i>General-purpose retrievers</i> | | | | | | | | | | | | |
| BM25 | 21.07 | 17.26 | 18.98 | 14.06 | 27.19 | 18.54 | 8.85 | 33.72 | 14.02 | 5.46 | 41.16 | 9.65 |
| DPR | 58.77 | 44.58 | 50.70 | 37.92 | 68.24 | 48.75 | 22.61 | 80.83 | 35.26 | 12.62 | 88.45 | 22.08 |
| Contriever | 29.69 | 23.47 | 26.22 | 19.14 | 35.67 | 24.91 | 12.19 | 44.27 | 19.11 | 7.31 | 52.12 | 12.82 |
| <i>Table-specific retrievers</i> | | | | | | | | | | | | |
| DTR | <u>49.40</u> | <u>37.77</u> | <u>42.81</u> | <u>32.88</u> | <u>59.60</u> | <u>42.38</u> | <u>21.04</u> | <u>74.50</u> | <u>32.81</u> | <u>12.37</u> | <u>86.22</u> | <u>21.63</u> |
| TableLlama | 47.99 | 36.93 | 41.74 | 32.70 | 59.48 | 42.20 | 20.34 | 72.44 | 31.76 | 11.55 | 81.56 | 20.24 |

Table 19: Multi-table retrieval results in terms of P/R/F1.

| Generator | Completeness | | | | | | | | | | | | | | |
|-------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | A&S | | | C&R | | | P&O | | | T&P | | | Avg. | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| <i>Proprietary LLMs</i> | | | | | | | | | | | | | | | |
| o3-mini | 67.40 | 62.92 | 62.84 | <u>63.78</u> | 57.55 | <u>58.78</u> | <u>65.65</u> | 57.69 | 59.77 | 65.28 | 57.33 | 59.42 | <u>65.53</u> | 58.87 | 60.20 |
| GPT-4o | 66.03 | 66.68 | 64.46 | 61.93 | 60.06 | 59.37 | <u>64.65</u> | 61.54 | 61.49 | 63.22 | 61.55 | 60.77 | <u>63.96</u> | 62.46 | 61.52 |
| Claude 3.5 Sonnet | 62.20 | <u>64.67</u> | 61.67 | 56.45 | 58.56 | 56.06 | 60.78 | 60.21 | 58.94 | 59.02 | <u>59.54</u> | 57.86 | 59.61 | <u>60.75</u> | 58.63 |
| <i>Open-source LLMs</i> | | | | | | | | | | | | | | | |
| DeepSeek-R1-8B | <u>68.61</u> | 64.24 | <u>64.23</u> | 62.21 | 58.15 | 58.72 | 65.29 | <u>60.89</u> | <u>61.41</u> | 62.92 | 59.02 | 59.48 | 64.76 | 60.58 | <u>60.96</u> |
| Qwen2-7B | 64.62 | 63.25 | 61.92 | 60.61 | <u>58.31</u> | 57.88 | 62.79 | 59.83 | 59.73 | 59.87 | 57.99 | 57.32 | 61.97 | 59.84 | 59.21 |
| Gemma-7B | 68.76 | 61.31 | 62.72 | 65.93 | 54.43 | 57.73 | 67.06 | 58.29 | 60.52 | <u>65.28</u> | 57.80 | 59.68 | 66.76 | 57.96 | 60.16 |
| Llama 3.1-8B | 59.24 | 63.43 | 59.20 | 56.21 | 57.16 | 55.11 | 58.95 | 59.44 | 57.45 | 57.02 | 58.07 | 55.85 | 57.85 | 59.52 | 56.90 |
| Mistral-7B | 66.62 | 63.69 | 62.96 | 62.16 | 57.26 | 58.00 | 63.53 | 59.13 | 59.66 | 62.05 | 57.80 | 58.29 | 63.59 | 59.47 | 59.73 |
| <i>SOTA TQA-methods</i> | | | | | | | | | | | | | | | |
| Dater | 63.53 | 59.97 | 59.70 | 60.57 | 56.96 | 57.26 | 64.53 | 59.23 | 60.35 | 62.20 | 59.12 | 59.16 | 62.71 | 58.82 | 59.12 |
| Chain-of-Table | 62.26 | 59.84 | 59.29 | 59.34 | 54.78 | 55.49 | 62.02 | 58.08 | 58.58 | 60.22 | 57.12 | 57.17 | 60.96 | 57.46 | 57.63 |
| TaPERA | 59.82 | 62.51 | 59.23 | 53.69 | 57.31 | 53.56 | 60.25 | 59.32 | 58.05 | 55.35 | 59.12 | 55.53 | 57.28 | 59.57 | 56.59 |

Table 20: MT-RAIG EVAL Completeness score results in terms of P/R/F1.

| Generator | SacreBLEU | ROUGE-L | METEOR | BERTScore | A3CU | TAPAS-Acc | G-Eval | |
|-------------------|-----------|---------|--------|-----------|-------|-----------|--------------|--------------|
| | | | | | | | Faithfulness | Completeness |
| Proprietary LLMs | | | | | | | | |
| o3-mini | 7.75 | 24.61 | 29.12 | 87.94 | 25.63 | 39.65 | 4.25 | 4.64 |
| GPT-4o | 19.16 | 33.92 | 38.91 | 90.02 | 35.87 | 34.15 | 4.34 | 4.69 |
| Claude 3.5 Sonnet | 15.59 | 30.11 | 35.99 | 89.40 | 31.87 | 31.77 | 4.28 | 4.51 |
| Open-source LLMs | | | | | | | | |
| DeepSeek-R1-8B | 12.07 | 30.03 | 28.48 | 89.22 | 31.32 | 54.70 | 3.85 | 4.25 |
| Qwen2-7B | 17.20 | 32.06 | 36.77 | 89.43 | 30.41 | 38.23 | 3.56 | 4.32 |
| Gemma-7B | 12.31 | 28.83 | 26.96 | 88.41 | 23.91 | 64.05 | 3.05 | 3.61 |
| Llama 3.1-8B | 9.55 | 23.85 | 35.62 | 86.45 | 27.03 | 61.25 | 3.57 | 4.14 |
| Mistral-7B | 15.07 | 30.20 | 34.66 | 89.00 | 27.51 | 43.12 | 3.55 | 4.30 |
| SOTA TQA-methods | | | | | | | | |
| Dater | 18.06 | 32.50 | 34.94 | 89.39 | 29.40 | 43.00 | 3.28 | 4.01 |
| Chain-of-Table | 18.15 | 32.38 | 35.60 | 87.33 | 33.26 | 42.08 | 3.69 | 4.25 |
| TaPERA | 13.89 | 28.57 | 32.74 | 88.32 | 24.61 | 44.67 | 2.63 | 4.00 |

Table 21: Insight generation results alongside existing automatic metrics. We report the average scores across all question types.

Question type : Analysis & Summary

Multi-table set Table 1

[TITLE] Spain in the Eurovision Song Contest 2009 - Semi-final 1 - Semi-final 1 - 14 February 2009

[HEADER] Draw | Artist | Song | Jury_votes | Televotes | Total_votes | Result_

[ROW 1] 1 | Yulia Valentayn | Uh la la | 6 | 3 | 9 | Out

[ROW 2] 2 | La Red de San Luis | Gracias, Madre Tierra | 8 | [b] | 2 | 10 | Out

[ROW 3] 3 | Vicente Casal | Tú me complementas | 1 | 7 | 8 | Out

...

Table 2

[TITLE] Spain in the Eurovision Song Contest 2009 - Semi-final 2 - Semi-final 2 - 21 February 2009

[HEADER] Draw | Artist | Song | Jury_votes | Televotes | Total_votes | Result_

[ROW 1] 1 | Diquesi | Subiré | 5 | 4 | 9 | Out

[ROW 2] 2 | Roel | Y ahora dices | 6 | 3 | 9 | Out

[ROW 3] 3 | Salva Ortega | Lujuria | 7 | 7 | 14 | Second Chance > Final

...

Table 3

[TITLE] Spain in the Eurovision Song Contest 2009 - Semi-final 3 - Semi-final 3 - 28 February 2009

[HEADER] Draw | Artist | Song | Jury_votes | Televotes | Total_votes | Result_

[ROW 1] 1 | El Secreto de Álex | Por esta vez | 7 | 6 | 13 | Out

[ROW 2] 2 | Beatriz | ¿Dónde te metes? | 4 | 1 | 5 | Out

[ROW 3] 3 | Remembrances | Espejismo | 3 | 4 | 7 | Out

...

Question

Summarize the performances of the artists who qualified for the final from each of the three semi-finals in Spain's Eurovision Song Contest 2009. What were their total votes, and how did they compare to those who did not qualify?

Insight

In Spain's Eurovision Song Contest 2009, the artists who qualified for the final from the three semi-finals showcased varying levels of performance based on their total votes. In the first semi-final, Noelia Cano received 11 votes, La La Love You garnered 20 votes, and Melody y Los Vivancos achieved 24 votes. Comparatively, those who did not qualify, such as Yulia Valentayn and La Red de San Luis, received 9 and 10 votes respectively, indicating that the qualifiers had a stronger reception. In the second semi-final, Soraya led with 24 votes, followed closely by Virginia with 20 votes and Jorge González with 16 votes, while non-qualifiers like Diquesi and Roel only managed 9 votes each, highlighting a clear distinction in popularity. The third semi-final saw Mirela, Isi, Santa Fe, and Julia Bermejo qualify with 18, 17, 24, and 14 votes respectively, while non-qualifiers like El Secreto de Álex and Beatriz received 13 and 5 votes, further emphasizing the gap in support. Overall, the artists who qualified consistently received higher votes than those who did not, reflecting their stronger performances and audience appeal across all semi-finals.

Question type : Comparison & Relationship

Multi-table set Table 1

[TITLE] car_1 - car_makers

[HEADER] Id | Maker | FullName | Country

[ROW 1] 1 | amc | American Motor Company | 1

[ROW 2] 2 | volkswagen | Volkswagen | 2

[ROW 3] 3 | bmw | BMW | 2

...

Table 2

[TITLE] car_1 - model_list

[HEADER] ModelId | Maker | Model

[ROW 1] 1 | 1 | amc

[ROW 2] 2 | 2 | audi

[ROW 3] 3 | 3 | bmw

...

Question

How does the id of a car maker relate to the specific models in the model list, and what does this indicate about the organization of car manufacturing data?

Insight

The id of a car maker is directly linked to the specific models in the model list, as each model is associated with a unique maker id that identifies the manufacturer responsible for it. For instance, the id 13 corresponds to Daimler Benz, which produces models like Mercedes-Benz and Mercedes, while id 19 is associated with Toyota, which includes models such as Scion and Toyota. This structured relationship indicates that car manufacturing data is organized in a way that allows for easy identification and categorization of models under their respective makers, facilitating data management and retrieval. Furthermore, the presence of makers like General Motors and Chrysler, which have designed more than three models, highlights how certain manufacturers have a broader portfolio, thus enriching the dataset and providing insights into the diversity of offerings from each maker. Overall, this organization enhances clarity and efficiency in understanding the connections between car makers and their models.

Table 22: MT-RAIG BENCH example for two question types, Analysis & Summary and Comparison & Relationship.

| Question type : Trend & Pattern | |
|---------------------------------------|---|
| Multi-table set | <p>Table 1
 [TITLE] La Cage aux Folles (musical) - 2004 Broadway revival - 2004 Broadway revival
 [HEADER] Year_ Award Category Nominee Result_
 [ROW 1] 2005 Tony Award Best Revival of a Musical Best Revival of a Musical Won
 [ROW 2] 2005 Tony Award Best Performance by a Leading Actor in a Musical Gary Beach Nominated
 [ROW 3] 2005 Tony Award Best Choreography Jerry Mitchell Won</p> <p>Table 2
 [TITLE] La Cage aux Folles (musical) - 2010 Broadway revival - 2010 Broadway revival
 [HEADER] Year_ Award Category Nominee Result_
 [ROW 1] 2010 Tony Award Best Revival of a Musical Best Revival of a Musical Won
 [ROW 2] 2010 Tony Award Best Performance by a Leading Actor in a Musical Kelsey Grammer Nominated
 [ROW 3] 2010 Tony Award Best Performance by a Leading Actor in a Musical Douglas Hodge Won</p> |
| Question | Compare the categories in which La Cage aux Folles was nominated for the Best Revival of a Musical in both the 2004 and 2010 Broadway revivals. What trends can be observed in the types of performances and productions recognized in these years? |
| Insight | In comparing the nominations for Best Revival of a Musical for La Cage aux Folles in the 2004 and 2010 Broadway revivals, several trends emerge regarding the types of performances and productions recognized. In 2004, the revival won the Tony Award for Best Revival and received nominations primarily focused on choreography, with Jerry Mitchell winning for Best Choreography, and a nomination for Best Performance by a Leading Actor in a Musical for Gary Beach. The emphasis was notably on choreography and costume design, reflecting a strong appreciation for the visual and performative aspects of the production. In contrast, the 2010 revival also won the Tony Award for Best Revival but expanded its recognition to include multiple categories, highlighting a broader range of performances with nominations for both Kelsey Grammer and Douglas Hodge in the Leading Actor category, as well as Robin de Jesus in the Featured Actor category. Additionally, the 2010 revival received nominations for Best Direction and various technical design categories, indicating a trend towards recognizing not only performances but also the overall production quality and direction. This shift suggests an evolving appreciation for comprehensive theatrical artistry, where both individual performances and the collaborative efforts of direction and design are increasingly acknowledged. |
| Question type : Performance & Outcome | |
| Multi-table set | <p>Table 1
 [TITLE] Elissa Sursara - Filmography - Filmography
 [HEADER] Year_ Film Role_ Language_ Notes
 [ROW 1] 2000 Hector Reno English Film
 [ROW 2] 2002 Marion Bridge Sam English Film
 [ROW 3] 2002 The Wet Season Taylor English Film
 ...</p> <p>Table 2
 [TITLE] Elissa Sursara - Television - Television
 [HEADER] Year_ Title Role_ Language_ Notes
 [ROW 1] 1994-96 Para Crancias Co-Host Portuguese 22 episodes
 [ROW 2] 1994-96 Touched by an Angel Natalie English The Heart of the Matter
 [ROW 3] 1996 American Gothic Alex English Requiem
 ...</p> |
| Question | What roles did Elissa Sursara undertake in the years 2002 and 2007 across her filmography and television appearances, and how do these roles reflect her career progression during that period? |
| Insight | In 2002, Elissa Sursara took on the roles of Sam in "Marion Bridge" and Taylor in "The Wet Season," alongside portraying Lena in the TV show "Taken," which positioned her primarily in supporting roles. By 2007, her filmography expanded significantly as she played Ava in "The Screening," Bena Brazilian in "Equal Opportunity," and Brittany in "Pain Within," while also taking on the character Jessica in the television series "Brothers & Sisters." This evolution from supporting roles in 2002 to more varied and prominent characters in 2007 reflects her career progression, highlighting her growing versatility and ability to engage with diverse genres and complex characters, marking a significant development in her acting career during that period. |

Table 23: MT-RAIG BENCH example for two question types, Trend & Pattern and Performance & Outcome.

| Criteria : Relevance | |
|-------------------------------|---|
| Multi-table set | <p>Table 1
 [TITLE] European Junior Swimming Championships Men Men
 [HEADER] Event Time_ Nationality Date_ Meet Location_
 [ROW 1] 50 m freestyle 22.06 Russia 14 July 2013 2013 Championships Poznan , Poland
 [ROW 2] 100 m freestyle 48.48 Russia 10 July 2009 2009 Championships Prague , Czech Republic
 [ROW 3] 200 m freestyle 1:46.58 France 18 July 2010 2010 Championships Helsinki , Finland
 ...
 Table 2
 [TITLE] FINA World Junior Swimming Championships Men Men
 [HEADER] Event Time_ Nationality Date_ Meet Location_

 [ROW 1] 50m freestyle 22.11 Australia 28 August 2013 2013 Championships Dubai , United Arab Emirates
 [ROW 2] 100m freestyle 48.97 United States 31 August 2013 2013 Championships Dubai , United Arab Emirates
 [ROW 3] 200m freestyle 1:47.55 Australia 27 August 2013 2013 Championships Dubai , United Arab Emirates
 ...
 Table 3
 [TITLE] List of Universiade records in swimming Men Men
 [HEADER] Event Time_ Nationality Date_ Meet Location_
 [ROW 1] 50m freestyle 21.67 Russia 16 July 2013 2013 Universiade Kazan , Russia
 [ROW 2] 100m freestyle 47.62 Russia 14 July 2013 2013 Universiade Kazan , Russia
 [ROW 3] 200m freestyle 1:44.87 Russia 15 July 2013 2013 Universiade Kazan , Russia
 ...</p> |
| Question | Analyze the significance of the locations where the European Junior Swimming Championships and the Universiade records in swimming are held. How do these venues contribute to the overall experience and performance of the athletes? |
| Verification rationale | The question does not relate to all the tables in the set, as it primarily mentions the European Junior Swimming Championships and Universiade records, while the other tables (such as the Hungarian records) are not considered. |
| Criteria : Faithfulness | |
| Multi-table set | <p>Table 1
 [TITLE] Aqeel Khan Singles runner-up Singles runner-up
 [HEADER] Date_ Tournament Surface Opponent_in_the_final Score
 [ROW 1] 28 June 2004 Gurgaon Hard Norikazu Sugiyama 6-1 3-6 4-6
 [ROW 2] 16 August 2004 Lahore Grass Toshiaki Sakai 4-6 6-7
 [ROW 3] 17 August 2006 Delhi Hard Yuri Bezeruk 6-7 2-6
 ...
 Table 2
 [TITLE] Aqeel Khan Singles titles Singles titles
 [HEADER] Date_ Tournament Surface Opponent_in_the_final Score
 [ROW 1] 15 August 2004 Islamabad Clay Toshiaki Sakai 7-6(3) 7-6(5)
 [ROW 2] 22 August 2004 Lahore Grass Toshiaki Sakai 1-6 6-4 6-3
 [ROW 3] 5 September 2004 Karachi Hard Tommaso Sanna 6-3 6-4
 ...</p> |
| Insight | Aqeel Khan faced Toshiaki Sakai in two finals, first on 15 August 2004 at the Islamabad tournament on Clay, where he won with a score of 7-6(3) 7-6(5), demonstrating strong performance and resilience. He then met Sakai again on 22 August 2004 at the Lahore tournament on Grass, where he lost 1-6 6-4 6-3, indicating his ability to recover after a poor start. Additionally, Khan faced Divij Sharan on 29 October 2007 at the Lahore tournament, also on Grass, where he lost 4-6 6-3 6-4, showcasing his competitive spirit as he managed to win the second set after losing the first. These matchups highlight Khan's competitive nature and adaptability, as he demonstrated the ability to perform well under pressure and recover from setbacks, even though he did not secure victories in all his finals. |
| Verification rationale | The answer does not cover all tables, focusing only on Table 2 (Aqeel Khan's singles titles) while omitting matches from Table 1 (Aqeel Khan's singles runner-up results). It also fails to fully integrate data from both tables. |

Table 24: Discarded data example in self-verification process for criteria about relevance and faithfulness.

| Criteria : Completeness | |
|-------------------------------|---|
| Question | Analyze the character arcs presented in the episodes directed by James Ormerod in Series 2: Spearhead in Germany and Series 3: Spearhead in Hong Kong. How do these arcs contribute to the overarching narrative of the Spearhead (TV series)? |
| Insight | The episodes directed by James Ormerod in Series 2 and Series 3 of Spearhead, specifically "New Brooms," "Matrimony," "Pressures," "Leave Takers," and "Repercussions" from Series 2, along with "Have a Happy Day" and "Another Happy Day" from Series 3, present character arcs that are integral to the series' overarching narrative, although specific details about these arcs are not provided in the available knowledge. The character development across these episodes likely reflects themes of personal growth, conflict resolution, and the impact of relationships, which are common in serialized storytelling. By exploring the dynamics between characters and their responses to various pressures and life changes, these arcs contribute to a deeper understanding of the characters' motivations and the series' exploration of societal issues. However, without explicit details on the character arcs, one can infer that the evolution of characters in these episodes serves to enhance the narrative by illustrating the complexities of human experiences in the context of the show's broader themes. |
| Verification rationale | The answer does not fully address the question, as it lacks specific details about the character arcs in the episodes directed by James Ormerod. While it touches on general themes and character development, it does not analyze the specific arcs mentioned, which are crucial for a complete response. |

Table 25: Discarded data example in self-verification process for criteria about completeness.

| MT-RAIG EVAL Faithfulness score : 47.37 | |
|--|--|
| Step : Table-aware insight decomposition | |
| Table schemas | <p>Table 1
[TITLE] 1988 Notre Dame Fighting Irish football team - Schedule - Schedule
[HEADER] Date_ Time_ Opponent_ Rank_ Result_ Attendance</p> <p>Table 2
[TITLE] 2005 Notre Dame Fighting Irish football team - Schedule - Schedule
[HEADER] Date_ Time_ Opponent_ Rank_ Site TV Result_ Attendance_</p> <p>Table 3
[TITLE] 1946 Notre Dame Fighting Irish football team - Schedule - Schedule
[HEADER] Date_ Opponent_ Rank_ Site Result_</p> <p>...</p> <p>Table 10
[TITLE] 1988 Minnesota Vikings season - Schedule - Schedule
[HEADER] Week Date_ Opponent Result_ Attendance</p> |
| Insight | The 1946 and 1988 Notre Dame Fighting Irish football teams both demonstrated impressive performances, but with notable differences in their ranking trajectories and strength of schedule. In 1946, Notre Dame started unranked but quickly ascended to #3 by their third game and maintained a #2 ranking for the majority of the season, facing only two ranked opponents. The 1988 team, however, began the season ranked #13 and steadily climbed to #1 by their ninth game, maintaining that top position for the remainder of the season. The 1988 schedule appears more challenging, with five ranked opponents, including a #1 Miami team, compared to the 1946 schedule. This suggests that the 1988 team faced a tougher slate of opponents, yet still managed to achieve and maintain the top ranking. The 1988 team's ability to consistently improve their ranking against stronger competition may indicate a more impressive overall season performance compared to the 1946 team, which maintained a high but steady ranking against fewer ranked opponents. |
| Decomposed claim set | <ol style="list-style-type: none"> 1. In 1946, Notre Dame started unranked but quickly ascended to #3 by their third game. 2. In 1946, Notre Dame maintained a #2 ranking for the majority of the season. 3. In 1946, Notre Dame faced only two ranked opponents. 4. The 1988 team began the season ranked #13. 5. The 1988 team steadily climbed to #1 by their ninth game. 6. The 1988 team maintained the top position for the remainder of the season. 7. The 1988 schedule included five ranked opponents. 8. The 1988 schedule included a #1 Miami team. 9. The 1946 schedule had fewer ranked opponents compared to the 1988 schedule. |
| Step : Claim verification | |
| Tables | <i>As referenced above</i> |
| Claim | In 1946, Notre Dame maintained a #2 ranking for the majority of the season. |
| Verification | [1.0, 1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 1.0, 0.0, 1.0, 1.0, 0.0, 0.0, 1.0, 1.0, 0.0, 1.0] |

Table 26: MT-RAIG EVAL Faithfulness evaluation example for each step based on Claude 3.5 Sonnet.

| | |
|--|--|
| MT-RAIG EVAL Completeness score : 69.57 | |
| Step : Question-aware insight decomposition | |
| Question | Summarize the attendance figures for the 1968 Buffalo Bills season and the 1969 Buffalo Bills season, focusing on the games against the Houston Oilers and Oakland Raiders. What might these figures indicate about fan engagement and support for the team during these years? |
| Ground-truth insight | The attendance figures for the 1968 Buffalo Bills season against the Houston Oilers were 34,339 on October 27 and 34,110 on December 7, while the games against the Oakland Raiders drew 43,056 on September 15 and 39,883 on November 28, indicating a moderate level of fan engagement despite the team's struggles that season. In contrast, the 1969 season saw increased attendance figures against the Houston Oilers, with 40,146 on September 21 and 46,485 on October 5, and a significant turnout of 54,418 for the game against the Oakland Raiders on October 19. These rising attendance numbers from 1968 to 1969 suggest a growing support and enthusiasm for the team, reflecting an improvement in fan engagement as the Bills likely began to show better performance or potential, which could have contributed to increased interest and attendance at their games. |
| Decomposed ground-truth topic set | <ol style="list-style-type: none"> 1. Attendance figures for 1968 Buffalo Bills season. 2. Attendance figures for 1969 Buffalo Bills season. 3. Attendance figures against Houston Oilers in 1968. 4. Attendance figures against Oakland Raiders in 1968. 5. Attendance figures against Houston Oilers in 1969. 6. Attendance figures against Oakland Raiders in 1969. 7. Moderate level of fan engagement in 1968. 8. Increased attendance figures in 1969. 9. Growing support for the team. 10. Improvement in fan engagement. 11. Team performance impact on attendance. 12. Enthusiasm for the team. |
| Predicted insight | In the 1968 Buffalo Bills season, the home game against the Houston Oilers on October 27 drew an attendance of 34,339, while the away game against the Oakland Raiders on November 28 had 39,883 attendees. In the 1969 season, the home game against the Houston Oilers on September 21 saw a slightly higher attendance of 40,146, and the away game against the Oakland Raiders on October 19 attracted 54,418 fans. These figures suggest that despite the Bills' poor performance in both seasons, with records of 1-12-1 in 1968 and 4-10 in 1969, there was a modest increase in attendance for these matchups, particularly in 1969. The higher attendance for the away games, especially against the Raiders, may reflect stronger fan engagement in cities with more competitive teams, while the relatively lower home attendance could indicate waning local support due to the Bills' struggles on the field. This trend highlights how team performance and the competitiveness of opponents can influence fan turnout and engagement. |
| Decomposed predicted topic set | <ol style="list-style-type: none"> 1. Attendance figures for 1968 Buffalo Bills season. 2. Attendance figures for 1969 Buffalo Bills season. 3. Attendance at home game against Houston Oilers. 4. Attendance at away game against Oakland Raiders. 5. Comparison of attendance figures between seasons. 6. Modest increase in attendance for matchups. 7. Higher attendance for away games. 8. Fan engagement in cities with competitive teams. 9. Lower home attendance for Buffalo Bills. 10. Impact of team performance on local support. 11. Influence of opponent competitiveness on fan turnout. |
| Step : Topic matching | |
| Ground-truth topic set | <i>As referenced above</i> |
| Predicted topic set | <i>As referenced above</i> |
| Matched ground-truth topic subset | [1, 2, 3, 4, 5, 6, 8, 11] |
| Matched predicted topic subset | [1, 2, 3, 4, 5, 6, 10, 11] |

Table 27: MT-RAIG EVAL Completeness evaluation example for each step based on GPT-4o.

| MT-RAIG EVAL Faithfulness : Claude 3.5 Sonnet (score: 60.53) v.s. Llama 3.1-8B (score: 24.32) | |
|---|--|
| Tables | <p>Table 1</p> <p>[TITLE] 1970 John Player League - Batting averages - Batting averages</p> <p>[HEADER] Player County Matches Innings Runs Average Highest_Score _100s _50s</p> <p>[ROW 1] Barry Richards Category:Articles with hCards Hampshire 12 11 592 65.77 155* 3 0</p> <p>[ROW 2] Clive Lloyd Category:Articles with hCards Lancashire 14 14 521 57.88 134* 1 2</p> <p>[ROW 3] Harry Pilling Category:Articles with hCards Lancashire 16 15 625 52.08 85 0 5</p> <p>...</p> <p>Table 2</p> <p>[TITLE] Somerset County Cricket Club in 2010 - Friends Provident t20 - Batting averages</p> <p>[HEADER] Player Matches Innings Runs Average Strike_rate Highest_Score _50s</p> <p>[ROW 1] James Hildreth Category:Articles with hCards 19 19 459 32.78 110.60 77* 2</p> <p>[ROW 2] Kieron Pollard Category:Articles with hCards 17 16 354 32.18 175.24 89* 2</p> <p>[ROW 3] Marcus Trescothick Category:Articles with hCards 19 19 572 31.77 157.14 83 6</p> <p>...</p> <p>Table 10</p> <p>[TITLE] Somerset County Cricket Club in 2009 - Match logs and statistics - Bowling averages</p> <p>[HEADER] Player Matches Overs Wickets Average Economy BBI _4wi</p> <p>[ROW 1] Ben Phillips Category:Articles with hCards 8 47.0 14 14.00 4.17 3/24 0</p> <p>[ROW 2] Charl Willoughby Category:Articles with hCards 6 42.4 11 19.18 4.94 3/36 0</p> <p>[ROW 3] Zander de Bruyn Category:Articles with hCards 8 32.4 6 28.33 5.20 4/20 1</p> <p>...</p> |
| Question | What can be inferred about the batting styles and strategies of players in the 1970 John Player League versus those in the Somerset County Cricket Club in 1891 based on their average scores and highest scores? |
| Claude 3.5 Sonnet | The data provided does not include information about the Somerset County Cricket Club in 1891, so a direct comparison between that year and the 1970 John Player League is not possible. However, we can draw insights from the 1970 John Player League data. In the 1970 season, batsmen like Barry Richards and Clive Lloyd demonstrated exceptional batting prowess, with Richards averaging an impressive 65.77 runs and Lloyd 57.88 runs. These high averages, coupled with their ability to score centuries (Richards had 3 and Lloyd had 1), suggest that the top batsmen of that era were capable of both consistency and explosive performances. The presence of multiple players with averages above 40 and the ability to score centuries or half-centuries indicates that the 1970 John Player League featured aggressive batting strategies, with players aiming for high individual scores while maintaining solid averages throughout the season. This style of play likely contributed to entertaining matches and high team totals, reflecting the evolving nature of limited-overs cricket in that period. |
| Llama 3.1-8B | The batting styles and strategies of players in the 1970 John Player League and the Somerset County Cricket Club in 2010 are distinct from those in the 1969 Player's County League and the Somerset County Cricket Club in 2009, as evidenced by their average scores and highest scores. In the 1970 John Player League, players like Barry Richards and Clive Lloyd achieved high average scores and highest scores, indicating a more aggressive batting approach. In contrast, players in the Somerset County Cricket Club in 2010, such as James Hildreth and Marcus Trescothick, had lower average scores and highest scores, suggesting a more defensive approach. This difference in batting style is also reflected in the 1969 Player's County League, where players like Keith Boyce and Peter Sainsbury had high average scores and highest scores, indicating an aggressive approach. However, the Somerset County Cricket Club in 2009, with players like Zander de Bruyn and Craig Kieswetter, had lower average scores and highest scores, suggesting a more defensive approach. This contrast highlights the evolution of batting styles and strategies over time, with teams and players adapting to different conditions and opponents. The Somerset County Cricket Club in 2010, with its more defensive ...Repeat similar contents... |
| Comparative analysis | Claude 3.5 Sonnet's response clearly establishes the premise that "a conclusion cannot be drawn due to the lack of data from 1891" and maintains consistency by focusing solely on the statistics within the 1970 John Player League. In contrast, Llama 3.1-8B's response expands the scope excessively by referencing multiple seasons and tournaments (1969, 2009, 2010, etc.), including information that is not directly relevant to the original question, which compares 1891 to 1970. As a result, while neither response reaches a perfect conclusion (due to retrieved tables), Claude 3.5 Sonnet's answer aligns more closely with the intent of the question by acknowledging the limitations of the available data and maintaining a coherent interpretation. |

Table 28: Comparison of MT-RAIG EVAL Faithfulness results between Claude 3.5 Sonnet and Llama 3.1-8B.

MT-RAIG EVAL Completeness : GPT-4o (score: 68.97) v.s. Qwen2-7B (score: 9.84)

| | |
|-----------------------------|---|
| Question | Summarize the performance of all Belarusian clubs in the European Cups across the 2006, 2007, and 2008 seasons, focusing on the rounds they reached and the teams they faced in each leg of the competition. |
| Insight | Across the 2006, 2007, and 2008 seasons, Belarusian clubs had varied performances in European Cups. In 2006, MTZ-RIPO Minsk participated in the UEFA Intertoto Cup, losing to Shakhter Karagandy in the first round and subsequently to Moscow in the second round. Dinamo Minsk and BATE Borisov also competed in the UEFA Cup, with Dinamo Minsk eliminated by Artmedia Bratislava and BATE Borisov losing to Rubin Kazan. Shakhtyor Soligorsk faced defeat in the UEFA Champions League against Široki Brijeg. In 2007, Shakhtyor Soligorsk advanced past Ararat Yerevan in the Intertoto Cup but was eliminated by Chernomorets Odessa in the second round. Dinamo Brest and Dinamo Minsk had mixed results in the UEFA Cup, with Dinamo Brest advancing past Liepājas Metalurgs but Dinamo Minsk failing against Odense. BATE Borisov faced Villarreal in the UEFA Cup and was eliminated, while in the Champions League, they had a narrow defeat against APOEL Nicosia but managed to progress past Hafnarfjarðar before losing to Steaua București. In 2008, Shakhtyor Soligorsk was eliminated in the Intertoto Cup by Cracovia, while Gomel and MTZ-RIPO Minsk both lost in the UEFA Cup qualifying rounds. BATE Borisov had a notable run in the Champions League, defeating Valur and reaching the group stage, where they faced tough competition from Real Madrid, Juventus, and Zenit St. Petersburg, ultimately finishing with mixed results. Overall, Belarusian clubs struggled to advance past the early rounds in most competitions during these seasons, with BATE Borisov achieving the most significant success by reaching the group stage of the Champions League in 2008. |
| GPT-4o | Across the 2006, 2007, and 2008 seasons, Belarusian clubs had mixed performances in European Cups, with BATE Borisov emerging as the most successful team. In 2006, BATE reached the second qualifying round of the UEFA Cup, defeating Nistru Otaci (3-0 aggregate) but losing to Rubin Kazan (5-0 aggregate). Shakhtyor Soligorsk exited in the first qualifying round of the UEFA Champions League against Široki Brijeg (0-2 aggregate), while MTZ-RIPO Minsk advanced past Shakhtyor Karagandy in the Intertoto Cup first round (6-4 aggregate) but fell to Moscow in the second round (3-0 aggregate). In 2007, BATE reached the third qualifying round of the UEFA Champions League, defeating APOEL Nicosia (3-2 aggregate) and Hafnarfjarðar (4-2 aggregate) before losing to Steaua București (4-2 aggregate). In the UEFA Cup, Dinamo Minsk defeated Skonto Riga (3-1 aggregate) in the first qualifying round but lost to Odense (5-1 aggregate) in the second, while Dinamo Brest was eliminated by Liepājas Metalurgs (3-2 aggregate) in the first qualifying round. Shakhtyor Soligorsk advanced to the second round of the Intertoto Cup, beating Ararat Yerevan (4-3 aggregate) but losing to Chernomorets Odessa (6-2 aggregate). In 2008, BATE had a historic run in the UEFA Champions League, progressing from the first qualifying round by defeating Valur (3-0 aggregate), then overcoming Anderlecht (4-3 aggregate) and Levski Sofia (2-1 aggregate) to reach the group stage, where they faced Real Madrid, Juventus, and Zenit St. Petersburg, earning two draws but failing to advance. In the UEFA Cup, MTZ-RIPO Minsk lost to Žilina (3-2 aggregate) in the first qualifying round, while Gomel was eliminated by Legia Warszawa (4-1 aggregate). Shakhtyor Soligorsk participated in the Intertoto Cup, defeating Cracovia (5-1 aggregate) in the first round but losing to Sturm Graz (2-0 aggregate) in the second. Overall, BATE's group stage appearance in 2008 marked a significant achievement for Belarusian football, while other clubs struggled to progress beyond the early rounds. |
| Qwen2-7B | In the European Cups across the 2006, 2007, and 2008 seasons, Belarusian clubs displayed a commendable performance, advancing to various stages of the tournaments. In 2006, BATE Borisov reached the UEFA Champions League group stage, defeating Anderlecht and Juventus, while Shakhtyor Soligorsk was eliminated in the UEFA Intertoto Cup's second round. In 2007, BATE Borisov made it to the UEFA Champions League group stage, overcoming APOEL Nicosia, Hafnarfjarðar, and Steaua București, and Shakhtyor Soligorsk advanced to the UEFA Intertoto Cup's second round. In 2008, BATE Borisov reached the UEFA Champions League group stage, facing Real Madrid, Juventus, and Zenit St. Petersburg, and Shakhtyor Soligorsk was eliminated in the UEFA Intertoto Cup's first round. These results highlight the competitive spirit and skill of Belarusian clubs in European football. |
| Comparative analysis | GPT-4o provides a thorough breakdown of how Belarusian clubs performed in European competitions during the 2006, 2007, and 2008 seasons, detailing the rounds they reached, the teams they faced, and the exact match results. In contrast, Qwen2-7B gives a much more general summary, primarily focusing on BATE Borisov while omitting other clubs like Shakhtyor Soligorsk and Dinamo Minsk, leaving out key details such as specific opponents, progression through different stages, and match outcomes. GPT-4o delivers a far richer response by fully covering each season's results and matchups, making it a much stronger answer. On the other hand, Qwen2-7B's response lacks depth due to its narrow focus on BATE and contains inaccuracies (e.g., claiming BATE defeated Juventus in 2006), resulting in a lower score. |

Table 29: Comparison of MT-RAIG EVAL Completeness results between GPT-4o and Qwen2-7B.