IS LLM-BASED TABULAR ANALYSIS INVARIANT TO LINGUISTIC REALIZATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) increasingly analyze tabular data through textual representations, allowing them to exploit semantic cues from column names, categorical values, and other natural-language elements. This capability, however, raises a fundamental question: whether LLMs can access and manipulate the same structured information consistently when its linguistic realization changes. We study this problem through a controlled English–Korean setting in which every query is fixed in English and semantically aligned tables are presented in either English or Korean. Our evaluation framework combines paired synthetic and real-world tables, diverse tabular analysis queries, multiple serialization formats, and paired metrics that distinguish aggregate performance differences from instance-level inconsistency. Across multiple LLM families, we find systematic but highly non-uniform sensitivity to linguistic mismatch. English tables generally yield higher accuracy, yet aggregate English–Korean performance gaps can substantially understate behavioral instability: models frequently change correctness across paired instances even when their average accuracy difference is small. Higher overall performance and larger models also do not consistently imply greater cross-lingual stability. Sensitivity varies sharply across query types and is most pronounced in cross-lingual linking, while remaining substantial even for basic column grounding. These results show that aggregate accuracy alone is insufficient to characterize whether LLMs consistently access semantically equivalent tabular information, highlighting the importance of paired cross-lingual evaluation for LLM-based tabular analysis.

1 INTRODUCTION

Large language models (LLMs) are increasingly being used to analyze tabular data by converting tables into textual representations and leveraging knowledge acquired during pretraining (Yin et al., 2020; Hegselmann et al., 2023; Liu et al., 2024; Nahid & Rafiei, 2024). Unlike conventional tabular models (Chen & Guestrin, 2016; Huang et al., 2020; Gorishniy et al., 2021; McElfresh et al., 2023; Lee et al., 2024; 2025) that primarily operate on numerical and categorical values, LLMs can directly interpret textual elements such as column names, categorical values, and natural-language descriptions. This capability provides an additional source of information, but also introduces a distinctive dependency: model behavior may depend not only on the structured information contained in a table, but also on *how that information is expressed linguistically*.

This dependency is particularly relevant for tabular data because the same underlying structure can admit multiple textual realizations (Zhang et al., 2023; Sui et al., 2024). A column may be described as *sex*, *feature 1*, or simply *X1*; similarly, a binary feature may be represented as *male/female*, *1/2*, or another pair of symbols. Such representations can preserve the same sample-wise structure and statistical relationships while exposing different semantic cues to an LLM. In the extreme, replacing meaningful feature names with abstract symbols removes semantic information without altering the underlying table itself. This raises a practical concern for LLM-based tabular analysis: *to what extent can model behavior be relied upon when the same structured information is expressed differently?*

The question is not simply whether English tables yield higher accuracy than tables expressed in another language. Given differences in language exposure during pretraining, some aggregate performance degradation under cross-lingual conditions may be unsurprising (Hu et al., 2020; Asai

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

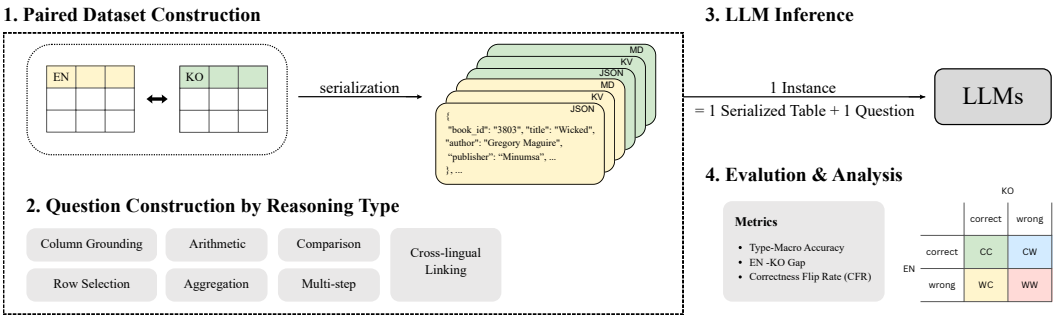


Figure 1: Overview of our controlled paired evaluation framework. Semantically aligned English–Korean tables preserve the same structured information while varying their linguistic realization. Using fixed English queries, we evaluate multiple tabular operations and serialization formats and compare language-matched and cross-lingual conditions through both aggregate performance and instance-level consistency.

et al., 2024; Xuan et al., 2025). More importantly, it remains unclear whether the same structured information is accessed consistently across linguistic realizations, whether stronger models are systematically more robust to such variation, and which tabular operations are most affected. These questions are particularly important because failures in basic access operations, such as identifying and grounding the relevant columns, can arise before more complex reasoning or prediction is performed.

In this work, we investigate linguistic mismatch as a controlled source of variation in LLM-based tabular analysis. As illustrated in Figure 1, we develop a paired evaluation framework that preserves the underlying table content, target semantics, and required reasoning operation while varying only the language in which the table is expressed. Every query is presented in English, whereas semantically aligned tables are presented in either English or Korean, with language-appropriate answer forms accepted for each table representation. This directly contrasts language-matched and cross-lingual access to the same structured information. Our evaluation spans seven fundamental tabular analysis operations, controlled synthetic and real-world data, multiple LLM families and parameter scales, and alternative table serialization formats.

Our results show that linguistic mismatch generally degrades aggregate performance, but its effect is not a simple uniform penalty. Models often change which instances they solve correctly across the two linguistic conditions, and substantial correctness flips can occur even when the aggregate English–Korean accuracy gap is small. Moreover, stronger overall performance or larger model size does not consistently imply greater cross-lingual stability. The instability also varies sharply across query types: it is most pronounced for tasks requiring stronger cross-lingual semantic linking, yet remains substantial even for basic column grounding. These patterns persist across real-world domains and serialization formats, indicating that aggregate accuracy alone can understate how strongly linguistic realization affects access to structured information.

Our main contributions are summarized as follows:

- We introduce the consistency of structured-information access under linguistic variation as a research question for LLM-based tabular analysis, focusing on whether semantically equivalent tables remain equally accessible when the table and query languages are matched or mismatched.
- We develop a controlled paired evaluation framework using semantically aligned English–Korean tables, spanning synthetic and real-world datasets, seven tabular analysis query types, multiple LLM families and parameter scales, and alternative serialization formats. The paired design explicitly separates aggregate performance differences from instance-level consistency.
- We show that linguistic mismatch induces systematic but highly non-uniform changes in model behavior. In particular, aggregate English–Korean accuracy gaps can substantially understate instance-level instability, while stronger overall performance and larger model size do not consistently correspond to greater robustness. The strongest effects occur for cross-lingual linking, but notable instability also appears in basic column grounding.

2 Related Work

LLMs for Tabular Learning. Prior work has shown that pretrained language models can exploit semantic information in tabular data, including column names, categorical labels, and associated text. TaBERT (Yin et al., 2020), HybridQA (Chen et al., 2020), and TAT-QA (Zhu et al., 2021) demonstrated semantic and numerical reasoning over tables, while more recent methods such as TabLLM (Hegselmann et al., 2023), TP-BERTa (Yan et al., 2024), and TabuLa-8B (Gardner et al., 2024) directly leverage textualized table representations for prediction and transfer. These studies highlight the value of textual semantics in tabular learning, but largely focus on English settings and do not explicitly examine whether semantically equivalent information remains accessible under different linguistic realizations.

Cross-Lingual Language Understanding. Multilingual language models support substantial cross-lingual transfer (Pires et al., 2019; Feng et al., 2022), yet performance can still depend on the languages used for queries and supporting evidence (Hu et al., 2020; Asai et al., 2024; Xuan et al., 2025). Query–evidence mismatch has therefore been widely studied in cross-lingual information retrieval and question answering (Amaral et al., 2008; Asai et al., 2021; Roy et al., 2020), where systems must access relevant information expressed in a different language from the query. Our work considers the analogous problem for structured data, asking whether LLMs can access and manipulate the same tabular information consistently when the table and query languages differ.

Multilingual Table Reasoning. Recent benchmarks such as MultiTAT (Zhang et al., 2025) and TableEval (Zhu et al., 2025) extend table reasoning to multilingual settings and report substantial variation across languages, but primarily compare aggregate performance across multilingual conditions. In contrast, we use paired, semantically aligned tables to isolate linguistic mismatch and measure both aggregate performance differences and instance-level consistency. We also evaluate multiple table serializations because prior work has shown that textual representation choices can affect LLM performance (Sui et al., 2024; Deng et al., 2024; Qiu et al., 2024).

3 Method

We develop a controlled paired evaluation framework to examine the effect of linguistic mismatch on LLM-based tabular analysis. For each table–question pair, the query is fixed in English, while the corresponding table is presented in either English or Korean. Across the two conditions, we preserve the underlying rows, columns, numerical values, categorical correspondences, target semantics, and required reasoning operation, varying only the linguistic realization of the table. Language-appropriate answer forms are accepted for each table representation. The English condition therefore represents a language-matched setting between the table and query, whereas the Korean condition requires cross-lingual access to semantically aligned structured information.

The framework combines controlled data construction, diverse tabular analysis queries, and paired evaluation. We consider seven query types—column grounding, row selection, aggregation, arithmetic, comparison, multi-step reasoning, and cross-lingual linking—to characterize where linguistic mismatch affects tabular analysis. Synthetic and real-world tables are used to balance control and practical relevance. We further evaluate Markdown, JSON, and key–value (KV) serializations as a robustness analysis to determine whether the observed linguistic sensitivity depends on a particular textual representation. Finally, our paired metrics capture both aggregate English–Korean performance differences and instance-level changes in correctness across the two linguistic conditions.

We construct semantically aligned English–Korean tables while preserving their underlying tabular content and structure. Textual elements, including column names and language-dependent categorical values, are transformed only when reliable semantic correspondence can be established. Numerical values and units are retained without conversion or localization, and attributes whose cross-lingual equivalence cannot be reliably verified are excluded. These controls are intended to minimize unintended changes beyond the linguistic realization of the table.

Our dataset combines synthetic and real-world tables for complementary purposes. Synthetic tables provide explicit control over vocabulary, table structure, and the semantic relationships required by each query, whereas real-world tables allow us to examine whether the observed effects persist

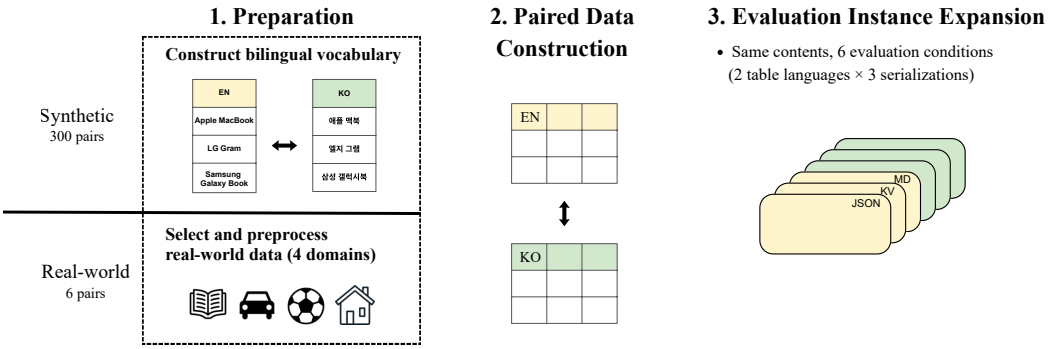


Figure 2: Detailed data construction pipeline for the paired English–Korean evaluation. Synthetic and real-world tables are constructed using predefined bilingual correspondences while preserving table structure, row alignment, numerical values, and task semantics across the two language versions. Each table–question pair is then expanded across two table languages and three serialization formats to form the final evaluation instances.

in more realistic data settings. In total, we construct 306 tables: 300 synthetic and 6 real-world, each instantiated in both English and Korean. From these tables, we derive 1,520 unique table–question pairs. Each pair is evaluated under two table languages and three serialization formats, yielding 9,120 evaluation instances: 6,900 synthetic and 2,220 real-world. Additional details on data construction, bilingual mappings, and representative examples are provided in Appendix A. The constructed evaluation dataset is provided as supplementary material.

3.0.1 SYNTHETIC DATA

We construct 300 synthetic tables, each instantiated as an English–Korean pair. Each table contains 30 rows and either six or eight columns. All tables are generated from predefined bilingual vocabularies with explicit English–Korean mappings, using a fixed random seed for reproducibility. Within each pair, the number, order, and underlying content of rows and columns are identical across the two language versions.

The synthetic tables are designed to isolate semantic relationships that are difficult to control in real-world data. We construct cases requiring alignment between query expressions and table headers, discrimination among semantically similar candidate attributes, linking of categorical values, and reasoning over multiple conditions. Accordingly, the bilingual vocabularies contain not only direct cross-lingual correspondences but also alternative expressions for the same concept and semantically related expressions that must be distinguished.

For example, categorical risk levels represented as *Low*, *Medium*, and *High* in English may correspond to *일반 관리 대상*, *우선 관리 대상*, and *집중 관리 대상* in Korean. These labels preserve the intended categorical distinctions while differing substantially in surface form. Likewise, entities such as *Samsung Galaxy Book*, *Apple MacBook*, and *LG Gram* are paired with *삼성 갤럭시북*, *애플 맥북*, and *엘지 그램*. Such constructions reduce reliance on direct string matching and require models to identify semantically corresponding attributes and values across linguistic realizations.

3.0.2 REAL-WORLD DATA

To examine whether the observed effects extend beyond controlled synthetic settings, we construct paired tables from four real-world domains: books (National Library of Korea, 2026), sports (Leone), automobiles (Cooper Union), and real estate (Kanchana, 2024). For each domain, we manually construct semantically aligned English–Korean counterparts from the source tables rather than relying on automatic translation. We use one paired table for each domain, except for books, which are divided into three subsets according to the correspondence between Korean and English titles: transliteration, literal translation, and free translation. This yields six real-world paired tables in total.

For each domain, we manually inspect the table attributes and retain only those for which reliable English–Korean correspondence can be established. Column headers and language-dependent cat-

Table 1: Overview of the seven query types used in our evaluation, together with representative examples and the corresponding reasoning operations. The query types are designed to probe complementary stages of tabular analysis, ranging from semantic grounding and row selection to numerical operations, multi-step reasoning, and cross-lingual semantic linking.

Type	**Description**	**Example Query**	**Expected Reasoning**
Column Grounding	Maps expressions in the query to the corresponding table attributes.	*Who wrote Demian?*	Map *wrote* to the *author* column, locate the row corresponding to *Demian*, and retrieve the author value.
Row Selection	Identifies rows satisfying one or more conditions.	*Which models are made by Ford?*	Identify the *make* column, select rows where *make* = *Ford*, and return the corresponding model values.
Aggregation	Aggregates values across multiple rows.	*What is the average number of bedrooms per property?*	Identify the *bedrooms* column, collect the relevant values, and compute their average.
Arithmetic	Performs arithmetic over retrieved numerical values.	*What is the wage difference between M. Dembele and M. Hummels?*	Retrieve the wage of each player and compute their difference.
Comparison	Compares values across rows or entities.	*Who is taller, R. Lewandowski or V. van Dijk?*	Retrieve the height of both players, compare the values, and return the taller player.
Multi-step Reasoning	Requires multiple operations to be composed sequentially.	*Which book has the highest borrow count per copy?*	Retrieve borrow count and number of copies for each book, compute the ratio, and identify the maximum.
Cross-lingual Linking	Links an English query concept to semantically corresponding Korean table values or categories whose surface forms differ across languages.	*Which defender has the highest wage?*	Link *defender* to the relevant Korean position categories, filter the corresponding players, and identify the maximum wage.

egorical values are translated using predefined mappings, while the same entities, row alignments, numerical information, and table structure are preserved across the paired versions. Numerical values, identifiers, units, and proper nouns are retained when no linguistic transformation is required. Long-form free-text attributes and other fields for which semantic equivalence cannot be reliably established are excluded. This conservative construction procedure is intended to minimize semantic discrepancies between the paired tables beyond their linguistic realization.

The book domain additionally provides a controlled range of lexical correspondence between Korean and English entity names. For all book pairs, we use the official titles published in each language rather than literal translations constructed for our experiments. For example, *데미안* and *Demian* primarily differ by transliteration, whereas *앵무새 죽이기* and *To Kill a Mockingbird* exhibit a more direct semantic correspondence across languages. In contrast, *천일야화* and *The Arabian Nights* differ substantially in surface form while referring to the same work. These subsets allow us to examine whether cross-lingual sensitivity changes as lexical correspondence between the two language realizations becomes weaker. The exact title mappings, grouping criteria, and publication sources used for verification are provided in Appendix A.

3.1 Query Types

We construct seven query types to probe distinct capabilities required for tabular analysis under linguistic mismatch, drawing on common reasoning operations in table question answering and structured reasoning (Glass et al., 2021; Zhu et al., 2021). Table 1 summarizes the definition, representative query, and expected reasoning process for each type. *Column grounding* examines whether expressions in an English query can be associated with the appropriate table attributes, while *row selection* evaluates whether records satisfying specified conditions can be identified. *Aggregation*,

arithmetic, and *comparison* test operations over retrieved values, and *multi-step reasoning* requires multiple operations to be composed sequentially.

We further include *cross-lingual linking* as a dedicated diagnostic category for cross-lingual semantic correspondence. These queries require the model to connect an expression in the English query with semantically corresponding table elements whose surface realizations differ in Korean. For example, a query referring to a *defender* may require linking that concept to the relevant position categories in the table before performing the requested operation. This category therefore directly tests cross-lingual semantic access in addition to the underlying tabular operation.

3.2 Evaluation Metrics

We evaluate model behavior from two complementary perspectives: aggregate performance and pair-level consistency. Aggregate metrics summarize performance across query types, while pair-level metrics measure whether correctness changes between semantically aligned English and Korean instances.

Aggregate Performance Metrics. Because the number of questions may differ across query types, particularly in the real-world datasets, we compute accuracy by assigning equal weight to each query type. Let T denote the number of query types and Acc_t the accuracy for type t. We define Type-Macro Accuracy as $\mathrm{MacroAcc} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{Acc}_t$.

We further quantify the directional performance difference between the language-matched and cross-lingual conditions as

$$\Delta\mathrm{MacroAcc} = \mathrm{MacroAcc}_{\mathrm{EN}} - \mathrm{MacroAcc}_{\mathrm{KO}} = \frac{1}{T}\sum_{t=1}^{T} \left(\mathrm{Acc}_{\mathrm{EN},t} - \mathrm{Acc}_{\mathrm{KO},t}\right).$$

A positive $\Delta\mathrm{MacroAcc}$ indicates higher performance when the table is presented in English, whereas a negative value indicates higher performance in the Korean-table condition.

Pair-Level Consistency Metrics. To enable a direct and controlled paired evaluation, we focus on two linguistic conditions, English and Korean. Because each table–question pair has semantically aligned versions in both languages, we additionally evaluate whether correctness is preserved across the two linguistic conditions. Each pair is assigned to one of four outcomes: CC if both versions are answered correctly, CW if only the English version is correct, WC if only the Korean version is correct, and WW if both are incorrect. We report the proportion of pairs belonging to each outcome category. Based on these outcomes, we define the Correctness Flip Rate (CFR) as $\mathrm{CFR} = \frac{N_{\mathrm{CW}}+N_{\mathrm{WC}}}{N}$, where N_{CW} and N_{WC} denote the numbers of CW and WC pairs, respectively, and N is the total number of English–Korean pairs. CFR therefore measures the proportion of instances whose correctness changes when the table language changes, regardless of direction.

The two metric groups capture complementary aspects of linguistic sensitivity. $\Delta\mathrm{MacroAcc}$ measures the directional difference in aggregate performance across query types, whereas CFR captures instance-level inconsistency. Consequently, a small aggregate performance gap may still coexist with a high rate of correctness changes across paired instances.

4 Results

Models. We evaluate six LLMs spanning multiple model families and parameter scales: GPT-4o-mini (OpenAI, 2024), GPT-5.4-mini (OpenAI, 2026), Llama-3.1-8B, Llama-3.2-3B (Dubey et al., 2024), Gemma-3-12B, and Gemma-3-27B (Team et al., 2025). The model set includes both proprietary and open-weight models and covers multiple model families and capacities. Qwen-3-8B was also considered but excluded because its long reasoning traces frequently exceeded the common token budget.

Prompting. We use zero-shot prompting with a fixed prompt template across all models and experimental conditions. The system prompt instructs each model to answer using only the provided

Table 2: Overall performance and pair-level consistency across models. EN and KO denote Type-Macro Accuracy (%) for English and Korean tables, respectively, and Gap denotes their difference in percentage points (pp). CW and WC indicate the proportions of pairs that change from correct to wrong and wrong to correct across the two language conditions, and CFR denotes the resulting Correctness Flip Rate.

	Type-Macro Accuracy (%)			Pair outcomes (%)		
Model	EN	KO	Gap	CW	WC	CFR
GPT-4o-mini	35.92	28.02	7.90	12.31	4.42	16.73
GPT-5.4-mini	45.93	36.40	9.53	16.40	6.87	23.27
Llama-3.1-8B	22.58	19.48	3.10	10.88	7.77	18.65
Llama-3.2-3B	16.68	11.86	4.82	9.07	4.25	13.31
Gemma-3-12B	30.10	19.17	10.93	15.26	4.30	19.56
Gemma-3-27B	35.42	26.44	8.98	12.92	3.94	16.86

table and to return a single short answer enclosed in `<answer>...</answer>`, without additional explanation. The user prompt contains the table language, serialization format, serialized table, and query. Complete prompt templates are provided in Appendix B.

4.1 Linguistic Mismatch Consistently Reduces Aggregate Accuracy

Table 2 reports overall Type-Macro Accuracy under the English and Korean table conditions. Across all six models, accuracy is lower when the fixed English query is paired with the semantically aligned Korean table. The English–Korean gap ranges from 3.10 to 10.93 percentage points (pp), indicating that the degradation occurs across different model families and parameter scales rather than being isolated to a particular model.

This difference emerges despite preserving the underlying table structure, values, target semantics, and required reasoning operation across each English–Korean pair. Linguistic realization alone can therefore materially affect access to the same structured information. The effect is also evident among the strongest-performing models: GPT-5.4-mini achieves the highest Type-Macro Accuracy in both conditions, yet decreases from 45.93% with English tables to 36.40% with Korean tables, a 9.53 pp drop—the second largest among the evaluated models despite its strongest overall performance. These results establish a consistent aggregate effect of linguistic mismatch, motivating a closer examination of whether the degradation reflects uniform performance loss or more complex changes in instance-level behavior.

4.2 Aggregate Performance Does Not Reveal Cross-Lingual Stability

Aggregate English–Korean accuracy differences do not fully reflect how sensitive a model is to linguistic variation. As shown in Table 2, Llama-3.1-8B exhibits the smallest English–Korean gap among the evaluated models at only 3.10 pp, yet 18.65% of paired instances change correctness across the two language conditions. Thus, similar aggregate performance can mask substantial changes in which individual instances are solved.

The same distinction appears even more clearly for GPT-5.4-mini. Despite achieving the highest Type-Macro Accuracy in both language conditions, it also exhibits the highest CFR at 23.27%. Its pair outcomes are strongly asymmetric: 16.40% of instances are correct only with English tables, whereas 6.87% are correct only with Korean tables. Higher overall performance therefore does not necessarily correspond to greater stability under linguistic mismatch.

Model size also shows a non-uniform relationship with robustness across the evaluated models. Within the Llama family, Llama-3.1-8B achieves higher accuracy and a smaller English–Korean gap than Llama-3.2-3B, but has a higher CFR (18.65% vs. 13.31%). In contrast, the larger Gemma-3-27B achieves higher accuracy while reducing both the aggregate gap and CFR relative to Gemma-3-12B. These contrasting patterns suggest that larger model size does not consistently correspond to greater cross-lingual consistency.

Table 3: English–Korean accuracy gaps and Correctness Flip Rates (CFRs) across models and query types. **Bold** values indicate the largest Gap or CFR within each query type.

	Column Grounding		Row Selection		Aggregation		Arithmetic		Comparison		Multi-step Reasoning		Cross-lingual Linking	
Model	Gap	CFR	Gap	CFR	Gap	CFR	Gap	CFR	Gap	CFR	Gap	CFR	Gap	CFR
GPT-4o-mini	1.81	11.26	4.66	14.00	2.67	8.00	0.40	10.80	8.00	22.67	**8.84**	15.51	28.89	34.88
GPT-5.4-mini	2.78	**26.86**	0.67	**15.33**	**5.33**	6.67	−8.67	**16.13**	4.00	20.00	8.08	**17.05**	**54.50**	**60.87**
Llama-3.1-8B	2.26	16.38	**8.00**	13.33	4.00	**9.33**	3.87	8.13	4.00	25.33	−0.90	16.79	0.48	41.26
Llama-3.2-3B	3.95	11.59	**8.00**	8.00	1.34	6.67	0.80	1.07	11.33	27.33	0.38	8.59	7.93	29.95
Gemma-3-12B	2.27	13.79	4.67	**15.33**	4.00	6.67	**7.34**	8.93	**38.66**	**44.00**	4.23	12.18	15.56	36.04
Gemma-3-27B	**4.60**	16.25	5.34	13.33	1.33	4.00	2.14	6.40	22.66	37.33	3.59	8.46	23.19	32.27

Together, these results show that aggregate accuracy, English–Korean performance gaps, and pair-level stability capture distinct aspects of model behavior. Evaluating linguistic robustness solely through average performance can therefore obscure substantial instance-level changes induced by the linguistic realization of the same structured information.

4.3 Linguistic sensitivity by reasoning type

Table 3 reveals substantial variation in linguistic sensitivity across query types. The strongest and most consistent instability appears in cross-lingual linking, where models must connect English query expressions to semantically corresponding table elements in Korean. GPT-5.4-mini exhibits a CFR of 60.87% and an accuracy gap of 54.50 pp for this query type, while several other models also show high flip rates. Comparison queries can likewise be highly sensitive, particularly for the Gemma models, although this pattern is less consistent across model families. These results show that the effect of linguistic mismatch depends strongly on the operation being performed.

More surprisingly, substantial instability also appears in basic semantic access. Column grounding requires the model primarily to associate an expression in the query with the appropriate table attribute, yet GPT-5.4-mini exhibits a CFR of 26.86% despite an accuracy gap of only 2.78 pp. Thus, linguistic mismatch can disrupt access to the relevant table semantics even before more complex numerical or multi-step reasoning is required.

The query-level results further show that linguistic mismatch does not induce a uniformly directional performance change. For Llama-3.1-8B, cross-lingual linking yields an English–Korean gap of only 0.48 pp despite a CFR of 41.26%, indicating substantial bidirectional changes in correctness that largely cancel at the aggregate level. In contrast, its multi-step reasoning gap is slightly negative at -0.90 pp, meaning that more paired instances are solved only in the Korean condition than only in the English condition. A larger negative gap appears for GPT-5.4-mini on arithmetic (-8.67 pp). These cases show that linguistic mismatch can produce not only large differences in instability, but also changes in the direction of the aggregate performance effect across models and query types.

Table 4 summarizes performance across the synthetic setting and real-world domains, while Table 5 further analyzes the book domain by translation type. Cross-lingual sensitivity varies across domains, with the largest average English–Korean gap observed in the book domain (10.90 pp), followed by sports (7.25 pp) and automobiles (5.63 pp). In contrast, real estate exhibits little aggregate difference between the two language conditions, with a slightly negative gap of -0.48 pp and a low CFR of 4.76%. These results indicate that the effect of linguistic mismatch can vary substantially across real-world data settings.

The book domain provides a more controlled view of this variation because its three subsets differ in the degree of correspondence between Korean and English titles. Both the aggregate gap and CFR increase as the title correspondence becomes less direct. The transliteration subset shows a gap of 8.89 pp and a CFR of 12.22%, while literal translation increases these values to 10.69 pp and 19.03%, respectively. The largest effect appears for free translation, where the gap reaches 17.08 pp and the CFR rises to 24.86%. This monotonic pattern suggests that greater cross-lingual sensitivity is associated with weaker lexical correspondence between semantically equivalent entity names.

Table 4: Average performance across synthetic and real-world data settings.

Data Setting	EN	KO	Gap	CFR
Synthetic	32.50	24.90	7.60	21.21
Book	45.34	34.44	10.90	18.70
Car	34.13	28.49	5.63	10.71
Real estate	12.46	12.94	−0.48	4.76
Sports	40.25	33.00	7.25	18.94

Table 5: Average performance across book subsets with increasing translation divergence.

Translation Type	EN	KO	Gap	CFR
Transliteration	49.86	40.97	8.89	12.22
Literal translation	50.42	39.72	10.69	19.03
Free translation	51.94	34.86	17.08	24.86

Importantly, the underlying entities and tabular structure remain fixed across the paired versions, indicating that the increased instability is associated with differences in surface realization rather than changes in the underlying structured information.

4.4 No Serialization Format Consistently Mitigates Linguistic Mismatch

Prior work has shown that table serialization can substantially affect LLM performance (Sui et al., 2024; Qiu et al., 2024). Table 6 shows, however, that no serialization format consistently improves robustness to linguistic mismatch. GPT-4o-mini and Llama-3.1-8B exhibit their smallest English–Korean gaps with the key–value representation, while GPT-5.4-mini performs most robustly with JSON. In contrast, Gemma-3-27B shows its smallest gap with Markdown. The relative ordering of Markdown, JSON, and KV therefore varies across models rather than revealing a universally robust representation.

Table 6: English–Korean Type-Macro Accuracy gap across serialization formats and models. **Bold** indicates the smallest gap for each model.

Model	Markdown	JSON	KV
GPT-4o-mini	9.39	8.33	**5.98**
GPT-5.4-mini	9.56	**7.28**	11.75
Llama-3.1-8B	3.90	2.97	**2.44**
Llama-3.2-3B	**3.95**	5.67	4.82
Gemma-3-12B	12.55	10.84	**9.48**
Gemma-3-27B	**8.44**	9.18	9.32

Moreover, the English–Korean gap remains positive across all three serialization formats for every model. Serialization can change absolute performance and the magnitude of the gap, but changing the textual representation alone does not eliminate linguistic sensitivity. These results suggest that cross-lingual robustness cannot be attributed to, or consistently improved by, a particular table serialization strategy.

5 Conclusion

We investigated linguistic mismatch in LLM-based tabular analysis using a controlled paired evaluation framework that preserves the underlying structured information while varying the table language between English and Korean. Across multiple models, query types, real-world settings, and serialization formats, we find that linguistic sensitivity is systematic but highly non-uniform: small aggregate English–Korean gaps can coexist with substantial instance-level correctness changes, stronger models are not consistently more robust, and sensitivity is most pronounced in cross-lingual linking while remaining substantial even for basic column grounding. These findings demonstrate that preserving the semantics of a table does not necessarily ensure consistent reasoning when the same information is linguistically realized in a different language. More broadly, they highlight a distinction between performance and robustness: aggregate accuracy alone is insufficient to characterize robustness to linguistic variation. Future multilingual evaluations of table understanding should therefore consider paired consistency alongside aggregate performance to capture whether model reasoning is preserved across languages. Our study is limited to an English–Korean setting with English queries, and extending this paired evaluation to additional languages and table–query language combinations remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, including improving clarity, grammar, and presentation of the manuscript. Generative AI was not used to generate experimental results, fabricate data, or make final scientific claims or conclusions. All AI-assisted text was reviewed and revised by the authors, and the underlying technical content, experimental design, analyses, and interpretations were independently verified by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or the collection of private or sensitive personal information. All real-world datasets used in our experiments are derived from publicly available data sources, and the synthetic datasets are generated solely for the purposes of controlled evaluation. We use these data only to study the behavior of LLMs in tabular analysis under linguistic variation. We are not aware of additional ethical concerns specific to this study beyond the general limitations and potential biases associated with publicly available datasets and pretrained language models.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the dataset construction procedure, query types, evaluation protocol, prompting setup, and evaluation metrics in the main paper and appendix. Additional information on the real-world data sources, preprocessing procedures, translation criteria, and prompt templates is also included in the appendix. The constructed evaluation dataset, including the paired English–Korean tables and corresponding queries, is provided as supplementary material.

REFERENCES

Carlos Amaral, Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, José Pina, and Cláudia Pinto. Priberam's question answering system in qa@ clef 2008. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 337–344. Springer, 2008.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. Xor qa: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 547–564, 2021.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1771–1800, 2024.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

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

Cooper Union. Car features and msrp. Kaggle. URL `https://www.kaggle.com/datasets/CooperUnion/cardataset`. Accessed: 2026-09-25.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 407–426, 2024.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 878–891, 2022.

Josh Gardner, Juan C Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling. *Advances in Neural Information Processing Systems*, 37:45155–45205, 2024.

Michael Glass, Mustafa Canim, Alfio Gliozzo, Saneem Chemmengath, Vishwajeet Kumar, Rishav Chakravarti, Avirup Sil, Feifei Pan, Samarth Bharadwaj, and Nicolas Rodolfo Fauceglia. Capturing row and column semantics in transformer based question answering over tables. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1212–1224, 2021.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems 34*, pp. 18932–18943, 2021.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pp. 5549–5581. PMLR, 2023.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pp. 4411–4421. PMLR, 2020.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Kanchana. Texas real estate trends 2024: 500 listings. Kaggle, 2024. URL `https://www.kaggle.com/datasets/kanchana1990/texas-real-estate-trends-2024-500-listings`. Accessed: 2026-09-25.

Kyungeun Lee, Ye Seul Sim, Hye-Seung Cho, Moonjung Eo, Suhee Yoon, Sanghyu Yoon, and Woohyung Lim. Binning as a pretext task: Improving self-supervised learning in tabular domains. *arXiv preprint arXiv:2405.07414*, 2024.

Kyungeun Lee, Moonjung Eo, Hye-Seung Cho, Dongmin Kim, Ye Seul Sim, Seoyoon Kim, Min-Kook Suh, and Woohyung Lim. MultiTab: A comprehensive benchmark suite for multi-dimensional evaluation in tabular domains. *arXiv preprint arXiv:2505.14312*, 2025.

Stefano Leone. Fifa 20 complete player dataset. Kaggle. URL `https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset`. The file players_19.csv was used. Accessed: 2026-09-25.

Tianyang Liu, Fei Wang, and Muhao Chen. Rethinking tabular data understanding with large 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)*, pp. 450–482, 2024.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C., Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems 36*, pp. 76336–76369, 2023.

Md Mahadi Hasan Nahid and Davood Rafiei. Tabsqlify: Enhancing reasoning capabilities of llms through table decomposition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5725–5737, 2024.

National Library of Korea. Library collection and loan data. Library Big Data Platform, 2026. URL https://www.data4library.kr/openDataV. Data retrieved from the May 2026 release.

OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, 2026. Accessed: 2026-09-24.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4996–5001, 2019.

Zipeng Qiu, You Peng, Guangxin He, Binhang Yuan, and Chen Wang. Tqa-bench: Evaluating llms for multi-table question answering with scalable context and symbolic extension. *arXiv e-prints*, pp. arXiv–2411, 2024.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. Lareqa: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 5919–5930, 2020.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532, 2025.

Jiahuan Yan, Bo Zheng, Hongxia Xu, Yiheng Zhu, Danny Chen, Jimeng Sun, Jian Wu, and Jintai Chen. Making pre-trained language models great on tabular prediction. In *International Conference on Learning Representations*, volume 2024, pp. 23570–23593, 2024.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8413–8426, 2020.

Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Shen Wang, Huzefa Rangwala, and George Karypis. Nameguess: Column name expansion for tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13276–13290, 2023.

Xuanliang Zhang, Dingzirui Wang, Keyan Xu, Qingfu Zhu, and Wanxiang Che. Multitat: Benchmarking multilingual table-and-text question answering. In *EMNLP (Findings)*, pp. 626–647, 2025.

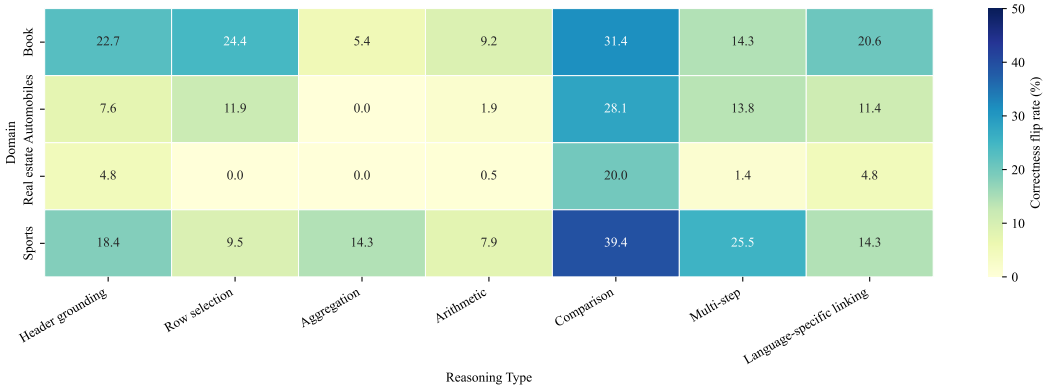
Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 3277–3287, 2021.

Junnan Zhu, Jingyi Wang, Bohan Yu, Xiaoyu Wu, Junbo Li, Lei Wang, and Nan Xu. Tableeval: A real-world benchmark for complex, multilingual, and multi-structured table question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7137–7157, 2025.

Domain	Source	Size	Retained attributes
Book (Transliteration)	Library Collection/Circulation Data (May 2026)	74×9	`book_id`, `title`, `author`, `publisher`, `year_of_publication`, `ISBN`, `classification_number`, `copies`, `borrows_count`
Book (Literal)	Same as above	74×9	Same as above
Book (Free)	Same as above	74×9	Same as above
Sports	FIFA 20 complete player	148×11	`fifa_id`, `name`, `age`, `height`, `weight`, `nationality`, `club`, `wage`, `preferred_foot`, `team_position`, `team_jersey_number`
Automobile	Car Features and MSRP	200×10	`make`, `model`, `year`, `engine_fuel_type`, `engine_HP`, `driven_wheels`, `number_of_doors`, `market_category`, `vehicle_size`, `vehicle_style`
Real estate	2024 Texas Real Estate	283×6	`list_price`, `baths`, `beds`, `sqft`, `type`, `year_built`

Table 7: Summary of the real-world datasets used to construct English–Korean table pairs.

Figure 3: Correctness Flip Rates(CFR) across real-world domains and reasoning types



A Real-world Data Construction

We construct real-world English–Korean table pairs from four domains: books, sports, automobiles, and real estate. Table 7 summarizes the source datasets, table dimensions, and retained attributes. Across all datasets, the paired versions preserve row-level correspondence and the underlying tabular information while varying language-dependent textual elements.

Book. The automobile data are derived from the May 2026 Library Collection and Loan Data released by the National Library of Korea (National Library of Korea, 2026). The original data are in Korean, and the corresponding English versions are constructed using officially published Korean and English editions whenever both are available.

The book data are divided into three subsets according to title translation complexity: *transliteration*, *literal translation*, and *free translation*. Transliteration primarily preserves the phonological form of the source title, literal translation directly renders its lexical meaning, and free translation

involves a larger degree of reformulation or adaptation between the source and target titles. These categories therefore reflect increasing degrees of cross-linguistic transformation between the officially published Korean and English title forms.

Each book is manually assigned to one of the three categories by comparing its officially published Korean and English titles. To improve the reliability of cross-language correspondence, we retain only books for which both Korean and English-language editions can be verified. Titles, author names, and other bibliographic expressions follow the forms used in the corresponding official publications whenever available. Each subset contains 74 rows and 9 columns.

Sports. The sports data are derived from `players_19.csv` file in the FIFA 20 complete player dataset (Leone). This data consist of 148 rows and 11 columns. The Korean version is constructed from the English source table. Numerical attributes, including age, height, weight, wage, and jersey number, are preserved without conversion.

Player names and club names are treated as proper nouns and are either retained in their original form or represented using established Korean forms commonly used in Korean sports broadcasts and news articles. Language-dependent categorical values, including nationality, preferred foot, and team position, are converted using predefined English–Korean mappings.

Automobile. The automobile data are derived from the Car Features and MSRP dataset and consist of 200 rows and 10 columns (Cooper Union). The Korean table is constructed from the English source table. Manufacturer and model names are preserved in their original form or represented using official names used by manufacturers or authorized sales channels.

Numerical attributes, such as model year, engine horsepower, and number of doors, are retained without conversion. Language-dependent categorical attributes, including fuel type, driven wheels, market category, vehicle size, and vehicle style, are translated using predefined one-to-one English–Korean mappings.

Real estate. The real-estate data are derived from the 2024 Texas real-estate dataset and consist of 283 rows and 6 columns (Kanchana, 2024). The retained attributes are `list_price`, `baths`, `beds`, `sqft`, `type`, and `year_built`. Because most retained cell values are numerical, the English–Korean transformation primarily affects column headers and the categorical property-type attribute. Numerical values are preserved without conversion or localization.

Attributes containing long-form natural-language content, such as `description`, are excluded during preprocessing because translation may introduce information addition, omission, or semantic alteration, making controlled correspondence across language versions difficult to maintain.

Cross-language alignment. Across all domains, we retain only attributes for which semantic equivalence across the two language versions can be explicitly verified. Within each pair, table dimensions, row ordering, corresponding entities, numerical values, identifiers, and units are preserved. Translation is restricted primarily to column headers and language-dependent categorical values. Categorical values are converted using predefined one-to-one mappings, while numerical values and units are retained without conversion or localization. These preprocessing rules are applied consistently to minimize unintended changes in the underlying tabular information across language versions.

B PROMPT TEMPLATE

We use the following fixed system and user prompts for all models and experimental conditions.

System Prompt.

```
You are a careful table question answering system.
Answer using only the provided table.
Return exactly one short final answer inside <answer>...</answer>.
Do not include explanation.
```

User Prompt Template.

```
Table language: {table_language}
Serialization: {serialization}
Table: {serialized_table}
Question: {question}

Return exactly one short final answer inside <answer>...</answer>.
```