Exploring the Robustness of Language Models for Tabular Question Answering via Attention Analysis

Anonymous authors
Paper under double-blind review

Abstract

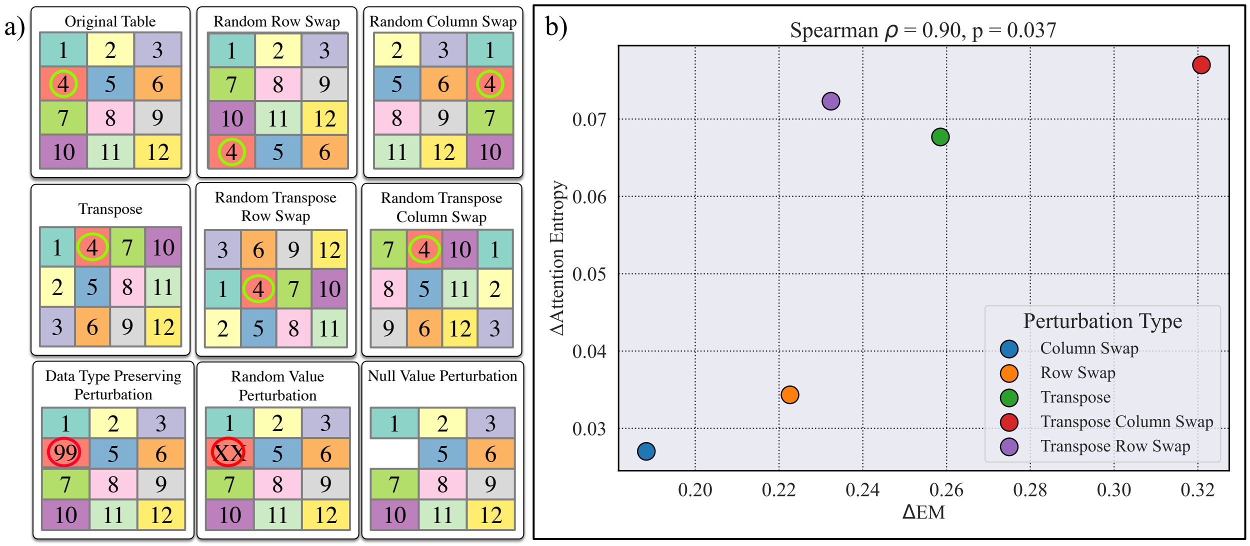
Large Language Models (LLMs), already shown to ace various unstructured text comprehension tasks, have also remarkably been shown to tackle table (structured) comprehension tasks without specific training. Building on earlier studies of LLMs for tabular tasks, we probe how in-context learning (ICL), model scale, instruction tuning, and domain bias affect Tabular QA (TQA) robustness by testing LLMs, under diverse augmentations and perturbations, on diverse domains: Wikipedia-based **WTQ**, financial **TAT-QA**, and scientific **SCITAB**. Although instruction tuning and larger, newer LLMs deliver stronger, more robust TQA performance, data contamination and reliability issues, especially on **WTQ**, remain unresolved. Through an in-depth attention analysis, we reveal a strong correlation between perturbation-induced shifts in attention dispersion and the drops in performance, with sensitivity peaking in the model's middle layers. We highlight the need for improved interpretable methodologies to develop more reliable LLMs for table comprehension. Through an in-depth attention analysis, we reveal a strong correlation between perturbation-induced shifts in attention dispersion and performance drops, with sensitivity peaking in the model's middle layers. Based on these findings, we argue for the development of structure-aware self-attention mechanisms and domain-adaptive processing techniques to improve the transparency, generalization, and real-world reliability of LLMs on tabular data.

1 Introduction

LLMs, despite being primarily trained on unstructured text, have demonstrated notable capabilities in structured data tasks, such as Tabular Question Answering (TQA). TQA requires models to interpret data presented in tables, demanding strong structural reasoning. TQA specifically challenges models to discern relationships and hierarchies implicit within tabular data, making it an ideal benchmark for structural reasoning capabilities. Assessing how LLMs navigate structured comprehension challenges can provide valuable insights into their robustness and reasoning capabilities (Borisov et al., 2023; Fang et al., 2024).

Recent studies emphasize the importance of robustness evaluations in understanding LLM behavior on structured tasks (Zhou et al., 2024). Specifically, perturbations in tabular structure and content significantly impact model performance, highlighting vulnerabilities that limit the practical reliability of models (Wang et al., 2022; Zhao et al., 2023). Furthermore, Liu et al. (2023) argues that LLMs inherently struggle with structural manipulations, advocating for an integrated approach combining symbolic reasoning to enhance model robustness.

Although these studies identify vulnerabilities and suggest potential improvements, there remains limited understanding of how internal model mechanisms respond to perturbations in structured data. Attention mechanisms form the core of transformer-based LLMs, governing how models distribute focus across input elements during processing (Clark et al., 2019). Prior work analyzing attention patterns in natural language tasks revealed that specific layers and attention heads critically influence model performance and robustness (Zhao et al., 2024; Barbero et al., 2025). These attention heads often serve specialized functions such as syntactic parsing or semantic alignment, highlighting the complexity of internal transformer mechanisms. However, detailed attention-level analyses for structured tasks, such as TQA, remain scarce.



Original Table

1	2	3
4	5	6
7	8	9
10	11	12

Random Row Swap

1	2	3
7	8	9
10	11	12
4	5	6

Random Column Swap

2	3	1
5	6	4
8	9	7
11	12	10

Transpose

1	4	7	10
2	5	8	11
3	6	9	12

Random Transpose Row Swap

3	6	9	12
1	4	7	10
2	5	8	11

Random Transpose Column Swap

7	4	10	1
8	5	11	2
9	6	12	3

Data Type Preserving Perturbation

1	2	3
99	5	6
7	8	9
10	11	12

Random Value Perturbation

1	2	3
XX	5	6
7	8	9
10	11	12

Null Value Perturbation

1	2	3
	5	6
7	8	9
10	11	12

Figure 1: (**a**) An example of the different possible table augmentation methods.(**b**) Scatter plot between change in attention entropy and EM score across perturbation types for `Llama-8B-Instruct` on **WTQ**.

To address this gap, we systematically investigate LLM robustness for TQA tasks across multiple dimensions, including *in-context learning*, *model scale*, *instruction tuning*, *domain biases*, *value perturbations*, and *model size*. Our experiments evaluate different perturbation types, highlighting how value-based alterations influence reasoning fidelity and faithfulness (Table 1). We also compare performance across diverse datasets to assess domain biases and generalizability.

Extending beyond surface-level performance metrics, we quantify changes in attention entropy across different attention heads and layers under various perturbations, analyzing how these changes correlate with performance degradation. Attention entropy effectively quantifies the dispersion in a model's attention distribution, providing a nuanced metric for understanding internal decision-making processes. With this, we aim to determine which attention heads are most sensitive to perturbations, leading to more substantial changes, and thus more distinct performance degradation.

Our study leverages diverse datasets: Wikipedia-based **WTQ** (Pasupat & Liang, 2015), financial report-based **TAT-QA** (Zhu et al., 2021), and scientific claims-based **SCITAB** (Lu et al., 2023). Each of these datasets provides a unique context and complexity, allowing us to assess the generalizability and domain-specific sensitivities of LLMs robustly. Through these diverse domains, our findings give insights into the sensitivity and reliability of LLMs on TQA, highlighting the broader challenge of understanding how LLMs reason over structured

2 Various Perturbation Categories

Each perturbation is designed to manipulate the table structure or content while preserving the inherent relational meaning of the table, thereby measuring robustness to table comprehension, as illustrated in Fig. 1.

2.1 Structural Perturbation (SP):

SP involves rearranging the columns and rows of the table to generate new examples. These perturbations simulate realistic scenarios where data presentation varies significantly, testing the model's ability to comprehend tabular structure. This ensures flexibility in understanding tabular data without distorting the semantics of the table. SP involves column swap, row swap, transpose, transpose column swap, and transpose row swap. These provide diverse perspectives on table comprehension, allowing for a thorough evaluation of the LLMs' ability to handle structural variations.

2.2 Value Perturbation (ValP):

ValP focuses on modifying the actual data values within tables, ensuring that the model accurately reflects the data. Value perturbations specifically challenge the semantic fidelity of the model by altering critical data points that directly affect the answer. We explore these types of **ValP**:

Data Type Preserving Perturbation (DVP): DVP involves altering the answer to the question and, respectively, the cell values within the table while maintaining their original data types. For instance, in Fig 9, given a question, “`What was the first venue for the Asian games?`”, we modify the correct answer, “`Bangkok, Thailand`", to “`Beijing`". These *counterfactual* entities test the faithfulness of the LLM to the tabular data. [1]. We utilize an automated counterfactual answer generation method that prompts GPT-3.5, ensuring the type correctness of the altered answer. Using a large language model, such fake answer generation makes it possible to generate fake answers that adhere to the data type and make the table semantically correct. Examples of prompts and details of DVP dataset generation are present in the Appendix A and B.

Random Value Perturbation (RVP): RVP(an example is shown in Fig. 10) relaxes DVP where instead of a counterfactual entity, we have a fixed string, e.g., “`r@nD0m v@1u3`”. Performance on this perturbation correlates with whether the injection of random data into the table affects the accuracy. The comparison of random and data type-preserving perturbation also highlights whether models are influenced by the injection of abstract values for table comprehension.

Null Value Perturbation (NVP): NVP removes the correct answer from the table completely. Evaluating the performance on NVP highlights the influence of Wikipedia content on solving the **WTQ** table question-answering task. LLMs that struggle more with the null value perturbation are likely to show consistent performance on TQA across different tabular datasets.

No Table (NT): To understand the extent of bias in **WTQ**, we evaluate LLMs on the no-table baseline. This approach further emphasizes the reliance of LLMs on Wikipedia content. By analyzing the performance of LLMs in the absence of the table, we can better understand the extent of dependence on the particular tabular data. This method helps to reveal the intrinsic capabilities of LLMs for TQA and their generalizability across different tabular datasets.

Collectively, these structural and value-based perturbations enable a comprehensive analysis of the model's ability to reason over table structure and content. By introducing such constrained and adversarial transformations, we can more precisely isolate the underlying factors that drive the performance in TQA.

3 Evaluation Metrics

Given the definition of different perturbations, we employ these three metrics to evaluate model performance under various perturbations in TQA tasks.

Exact Match Accuracy (EM): This metric calculates the proportion of instances for which the predicted answer exactly matches the ground truth answer. Formally, if N is the total number of instances, and $\text{correct}(i)$ is an indicator function that is 1 if the prediction for instance i is correct and 0 otherwise, then:

$$\text{EM} = \frac{\sum_{i=1}^{N} \text{correct}(i)}{N}.$$

Exact Match Difference (Emd)(Zhao et al., 2023; Zhou et al., 2024): Let EM_{orig} be the EM on the original (unperturbed) dataset, and $\text{EM}_{\text{perturbed}}$ be the EM after applying a perturbation. The Emd quantifies the change in EM due to perturbations:

$$\text{Emd} = \text{EM}_{\text{perturbed}} - \text{EM}_{\text{orig}}.$$

[1] We filter out the subset of data points where the table does not contain the answer, e.g., “`How many people stayed at least 3 years in office?`”

Negative values indicate performance degradation, while values close to zero imply robustness against perturbations.

Variation Percentage (VP)(Yang et al., 2022; Zhou et al., 2024): This metric measures the extent to which predictions change after applying perturbations. Given that, $C2W$ is the count of correct before perturbation and wrong after, and $W2C$ is the count of wrong before perturbation and correct after. Given N as the total number of instances, the variation percentage is:

$$\text{VP} = \frac{C2W + W2C}{N}.$$

A higher VP indicates greater sensitivity of predictions to perturbations, while a lower VP signifies more stable predictions.

In addition to performance-based metrics, we also examine internal model behavior through the analysis of attention patterns.

Attention Entropy: We analyze attention entropy to capture the dispersion of attention within different attention heads. Entropy serves as a measure for structural awareness, where high entropy indicates more evenly distributed attention, while low entropy reflects concentrated focus on a few tokens.

$$H_i = -\sum_j \mathbf{A}_{ij} \log(\mathbf{A}_{ij})$$

where $\mathbf{A}_{ij}$ is the attention weight to the j-th token.

4 Evaluation Performance

Operation	**WTQ Dataset**						**TAT-QA Dataset**						**SCITAB Dataset**					
	Large Model			**Small Model**			**Large Model**			**Small Model**			**Large Model**			**Small Model**		
	EM	VP	Emd	EM	VP	Emd	EM	VP	Emd	EM	VP	Emd	EM	VP	Emd	EM	VP	Emd
Original	0.39	-	-	0.29	-	-	0.36	-	-	0.23	-	-	0.117	-	-	0.138	-	-
Column	0.38	0.15	-0.01	0.28	0.15	-0.01	0.33	0.09	-0.03	0.19	0.10	-0.04	0.084	0.089	-0.032	0.137	0.093	-0.002
Row	0.32	0.18	-0.07	0.24	0.19	-0.05	0.32	0.11	-0.04	0.20	0.10	-0.03	0.098	0.083	-0.019	0.138	0.092	0.001
Transpose	0.37	0.17	-0.03	0.24	0.19	-0.05	0.34	0.09	-0.02	0.21	0.10	-0.02	0.094	0.087	-0.023	0.140	0.096	0.003
Transpose Row	0.35	0.19	-0.04	0.23	0.19	-0.06	0.29	0.13	-0.07	0.17	0.12	-0.06	0.103	0.088	-0.014	0.139	0.096	0.001
Transpose Col	0.27	0.23	-0.13	0.18	0.21	-0.11	0.31	0.11	-0.05	0.19	0.11	-0.04	0.099	0.087	-0.017	0.135	0.095	-0.002
NT	0.06	0.39	-0.33	0.02	0.29	-0.28	0.01	0.35	-0.34	0.00	0.23	-0.23	0.000	0.145	-0.145	0.000	0.138	-0.138
DVP	0.24	0.42	-0.14	0.17	0.36	-0.11	-	-	-	-	-	-	-	-	-	-	-	-
RVP	0.17	0.42	-0.21	0.13	0.35	-0.15	-	-	-	-	-	-	-	-	-	-	-	-
NVP	0.07	0.39	-0.31	0.03	0.29	-0.26	-	-	-	-	-	-	-	-	-	-	-	-

Table 1: The average EM, VP, and Emd of small and large LLMs on all perturbation operations. Large and Small LLMs are defined on Table 3.

4.1 Effects of ICL examples on TQA

Does instruction prompting assist LLM for better table comprehension for question-answering tasks? The heat maps in Figure 2 and Table 1 illustrate the performance of various LLMs, demonstrating that models fine-tuned for instructions or conversation exhibit improved performance. For instance, the `Llama3-70B-Instruct` model significantly outperforms its original version across all table transformations, indicating that instruction-based fine-tuning enhances the model's ability to handle complex reasoning tasks. Similarly, conversation-focused fine-tuning also yields better scores, albeit with a less pronounced improvement compared to instruction-focused tuning. This suggests that fine-tuning models on specific tasks like following instructions or conversing effectively enhances their capability to interpret and manipulate tabular data, making such approaches valuable for improving performance in structured data tasks. Table 1 also distinctly indicates that LLMs that have undergone instruction or conversation-based fine-tuning outperform their base counterparts in

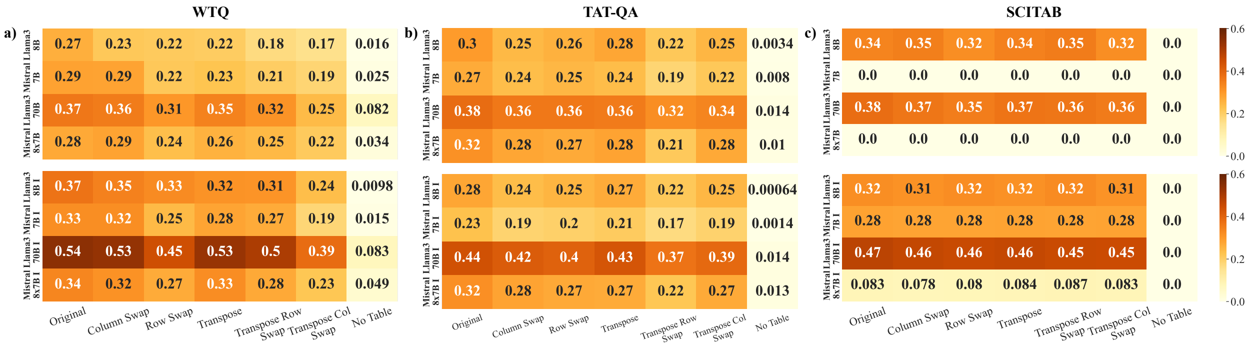


Figure 2: The average EM scores of the original models and instruction models(denoted by **I**) across various table augmentation techniques for **WTQ**, **TAT-QA** and **SCITAB** dataset with three fewshot examples.

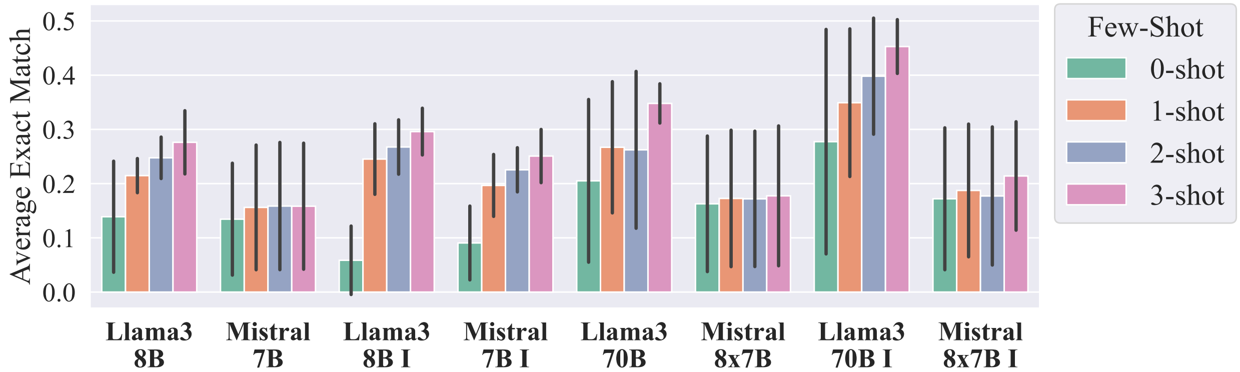


Figure 3: Comparison of average EM scores across models under varying few-shot settings (0-shot, 1-shot, 2-shot, and 3-shot). Here, **I** denotes the `Instruction` variant of the model.

TQA tasks for **SCITAB** dataset. Although **SCITAB** is inherently a classification task presented in a TQA format, using few-shot prompting provides valuable context, ultimately leading to more accurate and relevant responses.

4.2 Effects of the Model Type on TQA

Do newer models have better TQA abilities? Figure 4 shows that Llama3 models generally outperform the Mistral models across different configurations, indicating that newer architectures like Llama3 are more effective at table reasoning tasks. The 70B versions of these models generally perform better than their 7B variants, indicating that larger model sizes enhance reasoning capabilities. Overall, the advancements in model architecture and increased model size significantly contribute to better TQA abilities. Larger models (e.g., Llama3-70B, Mixtral-8x7B) generally show higher performance than smaller models (e.g., Llama3-8B, Mistral-7B), as seen in both the bar plot (Figure4) and heat maps (Figure2). For instance, Llama3-70B and Mixtral-8x7B have higher EM scores than Llama3-8B and Mistral-7B. This suggests that model size has a significant impact on TQA performance.

How do performances vary with value perturbations? [2] Table 1 shows the performance of various LLMs, as defined on Table 3, when subjected to different value-based perturbation on **WTQ** using different evaluation metrics: Exact Match(EM), Variation percentage (VP), and Exact Match Difference (EMD). We observe that LLMs experience a decline in EM scores across various operations compared to the original setup. The performance with RVP results in a significant performance drop, more so than DVP. This suggests a

[2] More details are in Appendices E

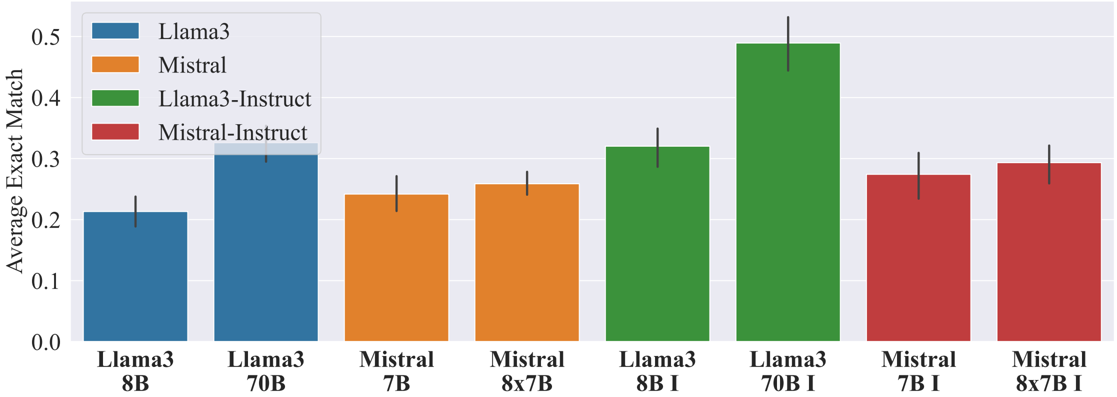


Figure 4: The average EM for different models across the **WTQ** dataset with 3-fewshot settings. The models are categorized by their types (Llama3, Mistral) and separated by their sizes (7B, 70B, etc.)

sharp decrease in the model's ability to process and comprehend tables when the insertion of arbitrary, non-contextual values compromises the table comprehension ability of LLMs. Conversely, the DVP result indicates that while the model struggles with content that deviates from the original data structure, maintaining data type consistency helps. VP increases significantly across all perturbations, indicating considerable changes in predictions due to the perturbations. EMD consistently shows negative values, with the most substantial performance drops occurring in the NT and NVP scenarios. The observed discrepancies in performance, particularly pronounced in DVP and RVP, underline a fundamental challenge: these models do not consistently apply their tabular comprehension capabilities when faced with perturbed tables.

4.3 Effects of Domain Specificity on TQA

How biased are LLMs towards Wiki-tables? Table 1 shows that when no table is provided, LLMs show a notable performance decline, emphasizing their dependence on tabular data to generate correct answers. Interestingly, the models still manage to answer about ≈5% of queries correctly, indicating a potential bias in Wiki-data. In the NVP scenario, where table values relevant to queries are nullified, there is a significant drop in performance, yet less severe compared to the complete absence of a table. This suggests that the models are biased by the contextual structured format cue even in the absence of relevant data.

How do out-of-box LLMs perform on specialized domains: TAT-QA and SCITAB? As shown in Table 1, LLMs exhibit moderate performance across various table augmentations on the **TAT-QA** dataset and **SCITAB** dataset. For **TAT-QA**, large models consistently outperform smaller models, underscoring their superior TQA abilities, but for **SCITAB**, the small models have better performance in comparison to larger models. The overall inconsistent scores still denote significant challenges inherent in niche domains, but instruction-tuned models are better here. From Table 1, the comparison among **WTQ**, **TAT-QA**, and **SCITAB** datasets reveals an interesting *accuracy-robustness tradeoff*: while models with higher EM on **WTQ** suffer larger EMD and higher VP, indicating higher sensitivity to perturbations, their relatively smaller EMD and lower VP on both **TAT-QA** and **SCITAB** reflect greater robustness, despite **SCITAB** 's overall lower baseline EM. Also notable is the sharp contrast in the NT performance, where on **TAT-QA** and **SCITAB** datasets, models have approximately 0% accuracy, a contrast from the **WTQ** performance. This suggests that performance on **WTQ** might be inflated due to biases favoring familiarity with Wikipedia, compared to niche domains like **TAT-QA** and **SCITAB**. This highlights the need for better benchmark design for tabular understanding. We also observe trends consistent with prior work: performance improves with few-shot prompting, instruction-tuned models generally outperform their base counterparts, and larger models tend to exhibit stronger TQA (Wei et al., 2022; Fang et al., 2024).

5 Attention Analysis

Understanding how structural perturbations in tabular inputs affect a model is critical for assessing robustness and diagnosing potential failure modes in TQA. As attention mechanisms regulate how language models allocate focus across table elements, examining their sensitivity to perturbations provides insights into representational stability. Here, we analyze this sensitivity by quantifying how perturbation-induced changes in attention dispersion correlate with performance degradation.

5.1 Effect of Perturbations on Attention Maps

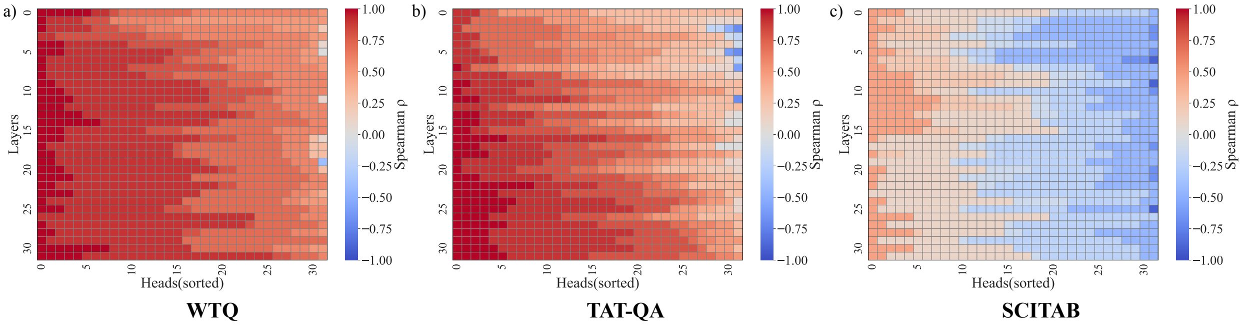
We examine how varying severities of structural perturbations influence attention weights using the `Llama3-8B-Instruct` model on the **WTQ** dataset. For each attention head across all layers, we measure the change in attention entropy between the original and perturbed table inputs. In parallel, we compute the corresponding change in Exact Match (EM) scores, capturing the performance impact of each perturbation.

Attention entropy captures the distribution uniformity of attention across tokens; higher entropy indicates diffuse attention, while lower entropy signifies focused attention. Thus, significant entropy changes reflect considerable shifts in the model's internal attentional focus due to perturbations.

Our analysis (Figure 1) reveals a significant positive correlation (Spearman $\rho = 0.90$, $p = 0.037$) between changes in attention entropy and EM degradation across perturbation types. This shows that more severe perturbations significantly disrupt the model's attention distribution, consequently diminishing its ability for TQA.

These findings highlight the sensitivity of attention mechanisms to structural integrity within tabular inputs. Perturbations affecting relational semantics induce greater attention dispersion and misalignment, directly impairing model accuracy. Thus, attention dispersion is a crucial indicator linking input perturbations with performance.

5.2 Analysis of Individual Attention Head Sensitivity to Perturbations



WTQ

TAT-QA

SCITAB

Figure 5: Heatmap showing Spearman correlation between changes in attention entropy and EM difference across all attention heads and layers in the `Llama3-8B-Instruct` model on the **a) WTQ**, **b) TAT-QA**, and **c) SCITAB** dataset

While averaging attention metrics across heads provides a general understanding of model behavior, prior research emphasizes that individual attention heads and specific layers, particularly the middle layers, significantly contribute to encoding task-relevant information (Barbero et al., 2025; Zhao et al., 2024). To investigate this at a finer granularity, we perform a layer-wise analysis of how perturbation-induced changes in attention entropy correlate with EM score degradation.

Specifically, we use the `Llama3-8B-Instruct` model on the **WTQ** dataset to compute the Spearman correlation between entropy changes (original vs. perturbed inputs) and corresponding EM differences for each attention head across all layers. Figure 5(**a**) demonstrates that correlations peak predominantly in the middle layers and remain consistently elevated through these central layers. This indicates that perturbation-induced shifts in attention distribution within middle layers are strongly predictive of performance degradation. The sharp correlation peak in **WTQ** suggests that the model relies heavily on mid-layer representations to align tabular structures with natural language queries, particularly when interpreting entity references and table schema. Distinct, though narrower, spikes at the input layer (0) and the output layers (30–31) further reveal that both the earliest token-encoding stage and the final representational consolidation phase are also vulnerable to structural perturbations.

Figure 5(**b**) and 5(**c**) extend the analysis to **TAT-QA** and **SCITAB**, respectively. The results reveal distinct patterns tied to domain characteristics. For **TAT-QA**, correlations are elevated not only in the middle layers but also extend into the upper layers of the model. In contrast, **SCITAB** exhibits a more sharply localized pattern, with peak correlations tightly concentrated within the middle layers. Although we observe a high correlation in these middle layers, the contrastive negative correlations are primarily due to the significantly poor EM performance on the **SCITAB** dataset. These findings reinforce the centrality of middle-layer attention mechanisms in tabular reasoning tasks, while also highlighting domain-specific variations in the vertical distribution of sensitivity. Additional experiments are included within the Appendix F for other models(`Llama3-8B`, `Mistral-7B`, and `Mistral-7B-Instruct`).

5.3 Correlation within Individual Attention

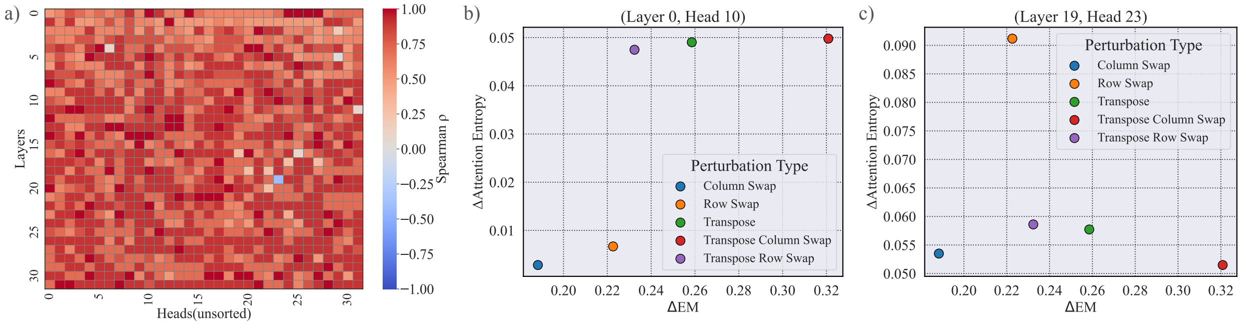


Figure 6: **(a)** Heatmap showing Spearman correlation between changes in attention entropy and EM difference across all attention heads and layers in the `Llama-8B-Instruct` model on the **WTQ** dataset, unsorted to depict actual correlation over each attention head. **(b-c)** Scatter plot showing the correlation between changes in attention entropy and EM scores across five structural perturbations for the **WTQ** dataset. **(b)** The plot highlights a strong positive correlation for Layer 0, Head 10, indicating high sensitivity of this head to structural perturbations. **(c)** In contrast, Layer 19, Head 23 shows the least correlation.

We find that structural perturbations in tabular inputs affect the internal attention dynamics of LLMs, specifically the `Llama3-8B-Instruct` model on **WTQ**, and that these dynamics are differentially expressed across the attention heads and layers. Figure 6(**a**) presents an unsorted Spearman correlation heatmap between per-head changes in attention entropy and the corresponding EM score differences, revealing substantial heterogeneity across all 32 layers and 32 heads. The heatmap captures a wide range of correlation strengths, from near-zero to values approaching $\rho = 0.9$, illustrating that not all attention mechanisms contribute equally to robustness. In particular, the concentration of high-correlation values in lower layers suggests that early-stage attention heads may play a foundational role in establishing reliable structural interpretations.

To unpack this variability, Figures 6(**b**) and 6(**c**) zoom in on two extreme heads identified from the heatmap. Layer 0, Head 10 (Figure 6(**b**)) exhibits a strong positive relationship: higher shifts in attention entropy due to structural perturbations reliably predict larger drops in EM scores. This implies that this head is critical for encoding spatial consistency or row-column alignment in tabular inputs, and disruptions to this alignment directly degrade model accuracy. The correlation is visually evident as a tight clustering of points around a positively sloped trend line across all five structural perturbation types (Row Swap, Column Swap, Transpose, Transpose Row Swap, and Transpose Column Swap).

In contrast, Layer 19, Head 23 (Figure 6(**c**)) demonstrates minimal correlation between entropy change and EM variation. This indicates a form of functional redundancy or robustness in this head's role; it either performs a task unrelated to structural parsing or maintains stable attention regardless of structural distortions. This wide range in sensitivity reinforces the notion that attention heads are functionally diverse and that only a subset contributes significantly to robustness under tabular perturbation.

Altogether, these findings suggest the possibility of identifying and selectively reinforcing robustness-critical attention heads through architectural tuning or fine-tuning objectives. By mapping correlation strengths across the entire model, we can isolate those mechanisms most affected by structural shifts and potentially develop strategies, such as attention regularization or selective re-weighting, that mitigate their susceptibility to disruption. This also opens avenues for pruning or interpretability studies focused on attention heads with minimal impact, offering insights into model compression or simplification without significant performance loss.

5.4 Cross-Model and Dataset Analysis of Perturbation Sensitivity

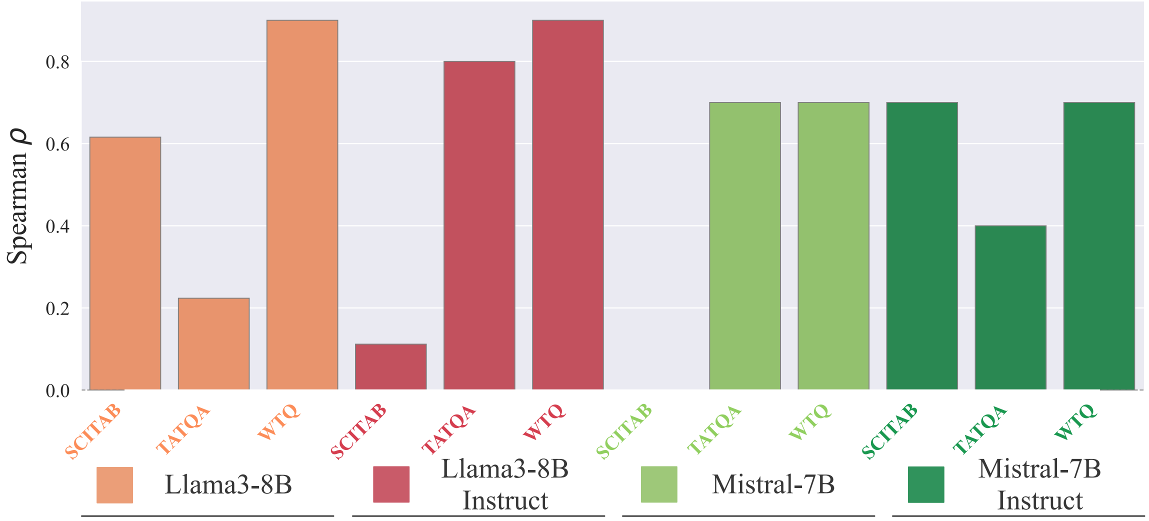


Figure 7: Spearman correlation between perturbation-induced attention-entropy change and EM degradation across 18 model–dataset pairs.

To evaluate the generality of our findings beyond the `Llama3-8B-Instruct` model on WTQ, we extend our correlation analysis between attention entropy change and EM performance degradation across a diverse set of model-dataset combinations. These include all the small models defined in 3, for which each models are tested on the **WTQ**, **TAT-QA**, and **SCITAB** datasets. For each configuration, we compute the Spearman correlation between the perturbation-induced changes in attention entropy and the corresponding drop in EM scores.

As shown in Figure 7, the positive correlation trend persists across nearly all settings. In particular, `Instruct` and `chat` variants of the models consistently show the strongest correlation values across datasets, reinforcing their heightened sensitivity to attention dispersion caused by structural perturbations. Notably, even non-instruct variants like `Mistral-7B` exhibit moderate to strong correlations, although the magnitude tends to be lower and more variable. We also find that **SCITAB** dataset results in most variation in the correlation, mainly due to the performance of the model, with the original table itself having a significantly low performance.

These consistent patterns across models and datasets confirm the robustness of our main claim: the severity of perturbation that induces greater shifts in attention entropy is reliably associated with a decline in model performance. This suggests that the change in attention entropy serves as a broadly applicable proxy for evaluating robustness in table-based QA models.

Moreover, this analysis demonstrates that the observed dynamics are not specific to any single dataset or model architecture. Instead, they reflect a more general representational vulnerability within current attention-based architectures when handling perturbed structured inputs.

6 Conclusion

Our study provides a comprehensive and robust analysis of LLMs under various perturbations for TQA. While larger, instruction-tuned models show improved performance, they remain highly sensitive to structural and value-based disruption. These disruptions notably manifest as perturbations such as random row swaps, column swaps, and transpositions, highlighting vulnerabilities in their structural reasoning capabilities. We also uncover domain biases, where models perform well on **WTQ** without tables but struggle on more specialized datasets, such as **TAT-QA** and **SCITAB**. Specifically, the performance on **WTQ** even in the absence of tables indicates reliance on memorized textual patterns from pretraining, rather than genuine tabular reasoning. In contrast, specialized domains such as financial and scientific tables pose more significant challenges due to their complexity and unique domain specificity.

We demonstrate that shifts in attention entropy, particularly in middle layers, are correlated with performance degradation. This observation was supported by detailed attention-level analyses, which revealed that attention heads in middle layers are particularly critical in encoding structural information, and their instability directly contributes to errors in comprehension. Furthermore, layer-wise attention analysis reveals that certain attention heads exhibit greater sensitivity, suggesting that targeted improvements in these areas could enhance robustness.

Our findings underline the critical need for improving both the interpretability and robustness of LLMs in TQA. Building on these insights, future research should focus on developing fine-grained attention-interpretability methods and domain-adaptive processing strategies to enhance transparency, cross-domain generalization, and reliability in practical applications.

7 Limitation

Although we provide extensive evaluation of LLMs on **WTQ**, **TAT-QA**, and **SCITAB** datasets, it is possible to include a broader range of datasets for a more comprehensive comparison that would highlight the generalizability of our method for both domain-specific datasets and Wikipedia-based datasets. While we anticipate that similar performance could be achieved with other tasks, such as table summarization, future work should include extensive analysis across various tasks and datasets to validate the assumption. Moreover, our study did not involve any structurally aware or fine-tuned models for tabular datasets. It is plausible that fine-tuning and structurally enhanced models could significantly impact the performance of different models. Additionally, our evaluation relied on exact match accuracy for assessing the text generation model's performance. This metric, while useful, limits the scope of evaluation for the question answering task. Future studies should employ more nuanced evaluation metrics to better assess the robustness of the models in TQA tasks. Moreover, we only conduct a case study of attention analysis with small models because of computation cost.

References

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličkovi ć, and Razvan Pascanu. Why do LLMs attend to the first token?, April 2025.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cEygmQNOeI.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding – A Survey, 2024. URL http://arxiv.org/abs/2402.17944.

Tianyang Liu, Fei Wang, and Muhao Chen. Rethinking Tabular Data Understanding with Large Language Models, 2023. URL http://arxiv.org/abs/2312.16702.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. SCITAB: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7787–7813, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.483.

Panupong Pasupat and Percy Liang. 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)*, pp. 1470–1480. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1142. URL http://aclweb.org/anthology/P15-1142.

Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. Robust (Controlled) Table-to-Text Generation with Structure-Aware Equivariance Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5037–5048. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.371. URL https://aclanthology.org/2022.naacl-main.371.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. TableFormer: Robust transformer modeling for table-text encoding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 528–537, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.40. URL https://aclanthology.org/2022.acl-long.40.

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. RobuT: A Systematic Study of Table QA Robustness Against Human-Annotated Adversarial Perturbations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6064–6081. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.334. URL https://aclanthology.org/2023.acl-long.334.

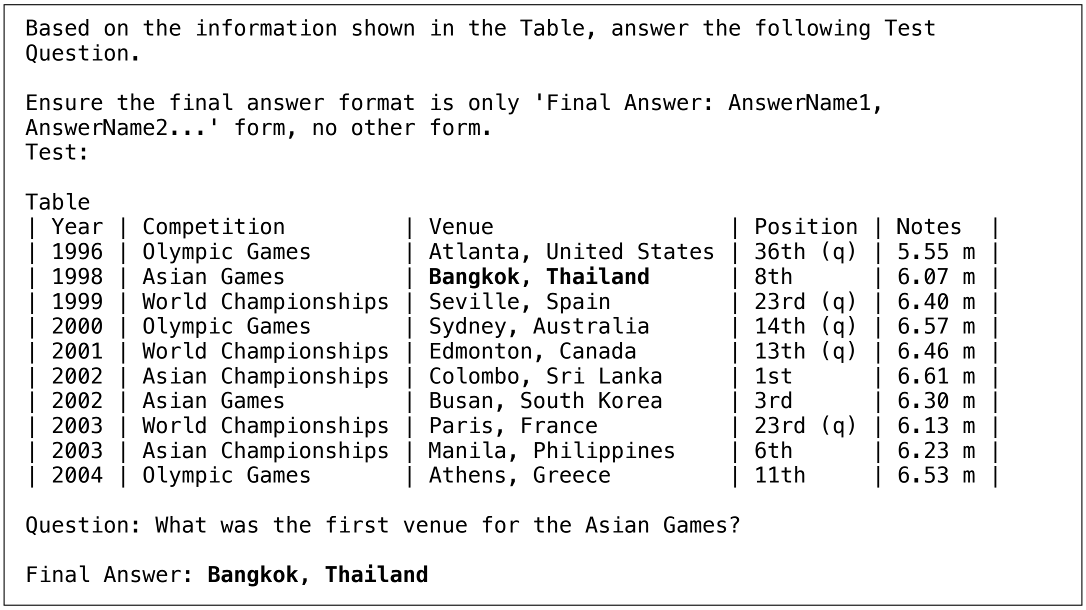
Zheng Zhao, Yftah Ziser, and Shay B Cohen. Layer by Layer: Uncovering Where Multi-Task Learning Happens in Instruction-Tuned Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15195–15214, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.847.

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. FREB-TQA: A fine-grained robustness evaluation benchmark for table question answering. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2479–2497, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.137. URL https://aclanthology.org/2024.naacl-long.137.

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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254.

A Example Prompt

Example of the prompts:



```
Based on the information shown in the Table, answer the following Test
Question.

Ensure the final answer format is only 'Final Answer: AnswerName1,
AnswerName2...' form, no other form.
Test:

Table
| Year | Competition          | Venue                   | Position | Notes  |
| 1996 | Olympic Games        | Atlanta, United States  | 36th (q) | 5.55 m |
| 1998 | Asian Games          | Bangkok, Thailand       | 8th      | 6.07 m |
| 1999 | World Championships  | Seville, Spain          | 23rd (q) | 6.40 m |
| 2000 | Olympic Games        | Sydney, Australia       | 14th (q) | 6.57 m |
| 2001 | World Championships  | Edmonton, Canada        | 13th (q) | 6.46 m |
| 2002 | Asian Championships  | Colombo, Sri Lanka      | 1st      | 6.61 m |
| 2002 | Asian Games          | Busan, South Korea      | 3rd      | 6.30 m |
| 2003 | World Championships  | Paris, France           | 23rd (q) | 6.13 m |
| 2003 | Asian Championships  | Manila, Philippines     | 6th      | 6.23 m |
| 2004 | Olympic Games        | Athens, Greece          | 11th     | 6.53 m |

Question: What was the first venue for the Asian Games?

Final Answer: Bangkok, Thailand
```

Figure 8: Example of a prompt with answer for **WTQ** dataset without Few Shot Prompt

B Generating Data Type Preserving Perturbation Dataset

We utilize ChatGPT-3.5 to generate type-preserving counterfactual answers using the prompt illustrated in Fig. 12. The prompt is designed to ensure that any generated answer aligns with the data type of the original answer while deliberately introducing a perturbation. For instance, given a question like "What was the first venue for the Asian games?" with the correct answer "Bangkok, Thailand," the model is prompted to output a different but type-consistent response, such as "Beijing." Similarly, for numerical data, a value like "42" might be replaced with another plausible number, such as "50," ensuring that the altered answer retains the original datatype constraints. This is achieved by explicitly defining the model's role as a generator of "fake" answers that preserve the data type of the original value. To ensure the validity of value perturbations, we select only tables that contain the exact answer to the given question. These constraints guarantee that data type preserving perturbations are applied exclusively to table-question-answer triples where the table explicitly contains the correct answer required to solve the question.

```
Based on the information shown in the Table, answer the following Test
Question.

Ensure the final answer format is only 'Final Answer: AnswerName1,
AnswerName2...' form, no other form.
Test:

Table
| Year | Competition         | Venue                   | Position | Notes  |
| 1996 | Olympic Games       | Atlanta, United States  | 36th (q) | 5.55 m |
| 1998 | Asian Games         | Beijing                 | 8th      | 6.07 m |
| 1999 | World Championships | Seville, Spain          | 23rd (q) | 6.40 m |
| 2000 | Olympic Games       | Sydney, Australia       | 14th (q) | 6.57 m |
| 2001 | World Championships | Edmonton, Canada        | 13th (q) | 6.46 m |
| 2002 | Asian Championships | Colombo, Sri Lanka      | 1st      | 6.61 m |
| 2002 | Asian Games         | Busan, South Korea      | 3rd      | 6.30 m |
| 2003 | World Championships | Paris, France           | 23rd (q) | 6.13 m |
| 2003 | Asian Championships | Manila, Philippines     | 6th      | 6.23 m |
| 2004 | Olympic Games       | Athens, Greece          | 11th     | 6.53 m |

Question: What was the first venue for the Asian Games?

Final Answer: Beijing
```

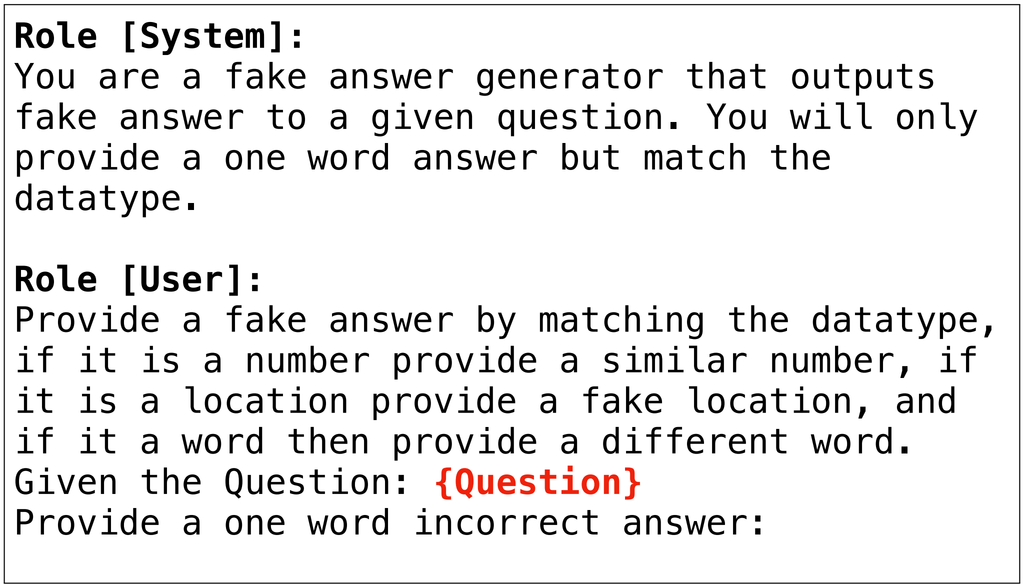
Figure 9: Example of a prompt with answer for **WTQ** dataset for Data Type Preserving Perturbation. In comparison to Fig. 8, we replace the correct answer(**Bangkok, Thailand**) with a fake answer(**Beijing**).

```
Based on the information shown in the Table, answer the following Test
Question.

Ensure the final answer format is only 'Final Answer: AnswerName1,
AnswerName2...' form, no other form.
Test:

Table
| Year | Competition         | Venue                   | Position | Notes  |
| 1996 | Olympic Games       | Atlanta, United States  | 36th (q) | 5.55 m |
| 1998 | Asian Games         | r@nD0m v@1u3            | 8th      | 6.07 m |
| 1999 | World Championships | Seville, Spain          | 23rd (q) | 6.40 m |
| 2000 | Olympic Games       | Sydney, Australia       | 14th (q) | 6.57 m |
| 2001 | World Championships | Edmonton, Canada        | 13th (q) | 6.46 m |
| 2002 | Asian Championships | Colombo, Sri Lanka      | 1st      | 6.61 m |
| 2002 | Asian Games         | Busan, South Korea      | 3rd      | 6.30 m |
| 2003 | World Championships | Paris, France           | 23rd (q) | 6.13 m |
| 2003 | Asian Championships | Manila, Philippines     | 6th      | 6.23 m |
| 2004 | Olympic Games       | Athens, Greece          | 11th     | 6.53 m |

Question: What was the first venue for the Asian Games?

Final Answer: r@nD0m v@1u3
```

Figure 10: Example of a prompt with answer for **WTQ** dataset for Random Value Perturbation. Here, we replace the correct answer(**Bangkok, Thailand**) with an abstract random value (**r@nD0m v@1u3**).

```
Based on the information shown in the Table, answer the following Test
Question.

Ensure the final answer format is only 'Final Answer: AnswerName1,
AnswerName2...' form, no other form.
Test:

Table
| Year | Competition          | Venue                   | Position | Notes  |
| 1996 | Olympic Games        | Atlanta, United States  | 36th (q) | 5.55 m |
| 1998 | Asian Games          |                         | 8th      | 6.07 m |
| 1999 | World Championships  | Seville, Spain          | 23rd (q) | 6.40 m |
| 2000 | Olympic Games        | Sydney, Australia       | 14th (q) | 6.57 m |
| 2001 | World Championships  | Edmonton, Canada        | 13th (q) | 6.46 m |
| 2002 | Asian Championships  | Colombo, Sri Lanka      | 1st      | 6.61 m |
| 2002 | Asian Games          | Busan, South Korea      | 3rd      | 6.30 m |
| 2003 | World Championships  | Paris, France           | 23rd (q) | 6.13 m |
| 2003 | Asian Championships  | Manila, Philippines     | 6th      | 6.23 m |
| 2004 | Olympic Games        | Athens, Greece          | 11th     | 6.53 m |

Question: What was the first venue for the Asian Games?

Final Answer: Bangkok, Thailand
```

Figure 11: Example of a prompt with answer for **WTQ** dataset for Random Value Perturbation. We remove the correct answer(**Bangkok, Thailand**).



```
Role [System]:
You are a fake answer generator that outputs
fake answer to a given question. You will only
provide a one word answer but match the
datatype.

Role [User]:
Provide a fake answer by matching the datatype,
if it is a number provide a similar number, if
it is a location provide a fake location, and
if it a word then provide a different word.
Given the Question: {Question}
Provide a one word incorrect answer:
```

Figure 12: Prompt designed for generating one-word, datatype-matching fake answers to questions from the **WTQ** dataset.

C Evaluation Dataset Size

We select three different datasets for comparison, **WTQ**, **TAT-QA**, and **SCITAB** datasets, with different numbers of evaluation datasets as described in Table 2. Each dataset provides distinct characteristics, **WTQ** is composed of Wikipedia tables, **TAT-QA** centers on financial reports, and **SCITAB** focuses on scientific claims, which together enable a comprehensive and robust assessment. For fair comparison, we limit the number of cell elements (< 150) within the table for both datasets. This constraint ensures that models are not disproportionately affected by excessively large tables, which could skew performance due to context window limitations. Moreover, keeping table size bounded allows for more consistent measurement of the effects of perturbations across datasets.

Similarly, for Value Perturbation, some queries relate to the overall structure of the table. Hence, we filter only those tables that contain the answer value for the given query. This filtering ensures semantic alignment between the question and the modified table, avoiding misleading evaluations where no correct answer exists in the perturbed version. In the case of the **WTQ** dataset, for instance, not all table-question pairs are amenable to value alterations, particularly when the table structure is insufficiently informative or lacks direct answer candidates. Altogether, this preprocessing pipeline yields a curated benchmark that is well-suited for analyzing perturbation sensitivity while controlling for confounding factors related to table size and answerability.

Operation	**Number of Pairs**
WTQ dataset	
Row Swap	204
Column Swap	204
Transpose	204
Transpose Row Swap	204
Transpose Column Swap	204
Data Type Preserving	141
Random Value	141
Null Value	141
No Table	204
TAT-QA dataset	
Row Swap	1668
Column Swap	1668
Transpose	1668
Transpose Row Swap	1668
Transpose Column Swap	1668
No Table	1668
SCITAB dataset	
Row Swap	1225
Column Swap	1225
Transpose	1225
Transpose Row Swap	1225
Transpose Column Swap	1225
No Table	1225

Table 2: Size of the Evaluation datasize

D Models

We selected recent open-source models that have been extensively studied and analyzed. Table 3 lists all the models we considered with their parameter size and their date of release. We include both base and instruction-tuned variants, allowing us to explore not only the effect of scale but also the impact of task specialization on tabular reasoning. The models span three major families: `Llama-2`, `Llama-3`, and `Mistral`, which together represent some of the most widely adopted transformer architectures in the open-source ecosystem. The inclusion of instruction-tuned variants is particularly important, as these models are optimized for following natural language instructions. This ability has been shown to influence performance on tasks that require a structured understanding significantly.

Moreover, by categorizing the models into 'small' and 'large' based on parameter count, we aim to systematically assess how model scale interacts with robustness and accuracy under various perturbation regimes. Recent models, such as `Llama-3` and `Mistral`, demonstrate architectural innovations, including improved token representations and mixture-of-experts routing, which provide a richer set of inductive biases for our evaluation. Such a comprehensive suite enables a detailed comparison of robustness across architecture, scale, and fine-tuning strategies, thereby supporting more generalizable insights into LLM behavior on TQA tasks.

Model	**Size**	**Date Released**
Small Model(Less than 10B parameter)		
`Llama-2-7b-hf`	6.74B	July 2023
`Llama-2-7b-chat-hf`	6.74B	July 2023
`Mistral-7B-v0.1`	7.24B	Sept 2023
`Mistral-7B-Instruct-v0.1`	7.24B	Sept 2023
`Meta-Llama-3-8B`	8.03B	April 2024
`Meta-Llama-3-8B-Instruct`	8.03B	April 2024
Large Model(Larger than 40B parameter)		
`Llama-2-70b-hf`	69B	July 2023
`Llama-2-70b-chat-hf`	69B	July 2023
`Mixtral-8x7B-v0.1`	46.7B	Dec 2023
`Mixtral-8x7B-Instruct-v0.1`	46.7B	Dec 2023
`Meta-Llama-3-70B`	70.6B	April 2024
`Meta-Llama-3-70B-Instruct`	70.6B	April 2024

Table 3: All the models with their parameter size and their date released. Large Models are defined as models with parameters larger than 40 billion parameters, and Small Models are models with parameters smaller than 10 billion parameters.

E Performance over Table Complexity

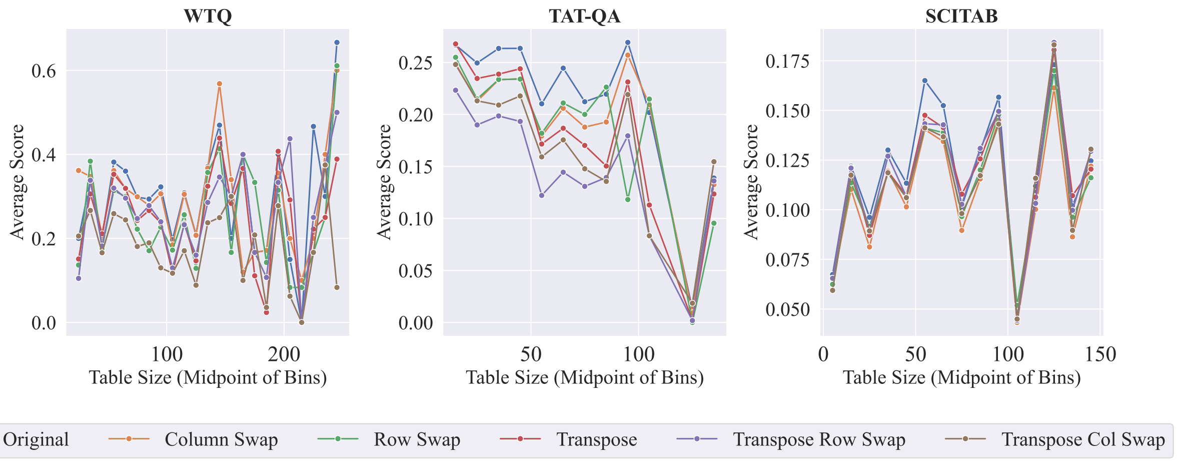


Figure 13: The average EM score for the different structural perturbations over the different sizes of tables on **WTQ**, **TAT-QA**, and **SCITAB** datasets

Figure 13 presents average model performance across varying table sizes for three tabular question-answering datasets—**WTQ**, **TAT-QA**, and **SCITAB**—under several table structure perturbations. Each plot compares the original condition with different table augmentation operations, such as column swapping, row swapping, and transpose-based modifications.

For **WTQ**, performance is low overall and highly variable across increasing table sizes, with no single condition consistently outperforming others. In contrast, **TAT-QA** shows a generally higher and more stable baseline, though still affected by growing table size and perturbations; performance tends to decline as tables grow larger, suggesting sensitivity to complexity. **SCITAB** results are somewhat mixed, with fluctuations at different size intervals, but the performance remains closer among the different conditions.

Across all three datasets, these results highlight that large language models, though capable, display uneven robustness to structural manipulations of tables. Although instruction tuning and larger model scales improve performance, structural changes continue to pose challenges, underscoring the need for more robust, structure-aware approaches to ensure reliable table comprehension.

F Attention Matrix Analysis

F.1 Spearman Correlation within All Attention Heads

Across both the model families `Llama3` and `Mistral`, the middle layers consistently exhibit the strongest positive Spearman correlations between perturbation-induced changes in attention entropy and EM degradation. This recurring 'hot spot' in the mid-layer shows a general architectural property; middle transformer blocks appear to serve as critical junctions where structural perturbations most directly translate into downstream performance loss. Such robustness vulnerabilities likely stem from these layers' dual role in integrating lower-level token interactions and preparing higher-level semantic abstractions, making them both information-rich and sensitive to distributional shifts.

However, differences emerge when contrasting base versus chat- or instruction-tuned variants. In the base models (`Llama3-8B`, `Mistral-7B`; Figure (14, and 15) subfigures **a-c**), the correlation patterns outside the middle layers fluctuate markedly, with a mixture of weak or even negative correlation, especially apparent on **TAT-QA** and **SCITAB** tasks. Specifically, for Figure 15(**c**),`Mistral-7B` does not correlate because for all the questions EM was 0. These oscillations align with the base models' generally lower EM scores on these datasets, suggesting that noisy or unreliable predictions can obscure the coherent relationships between entropy and performance.

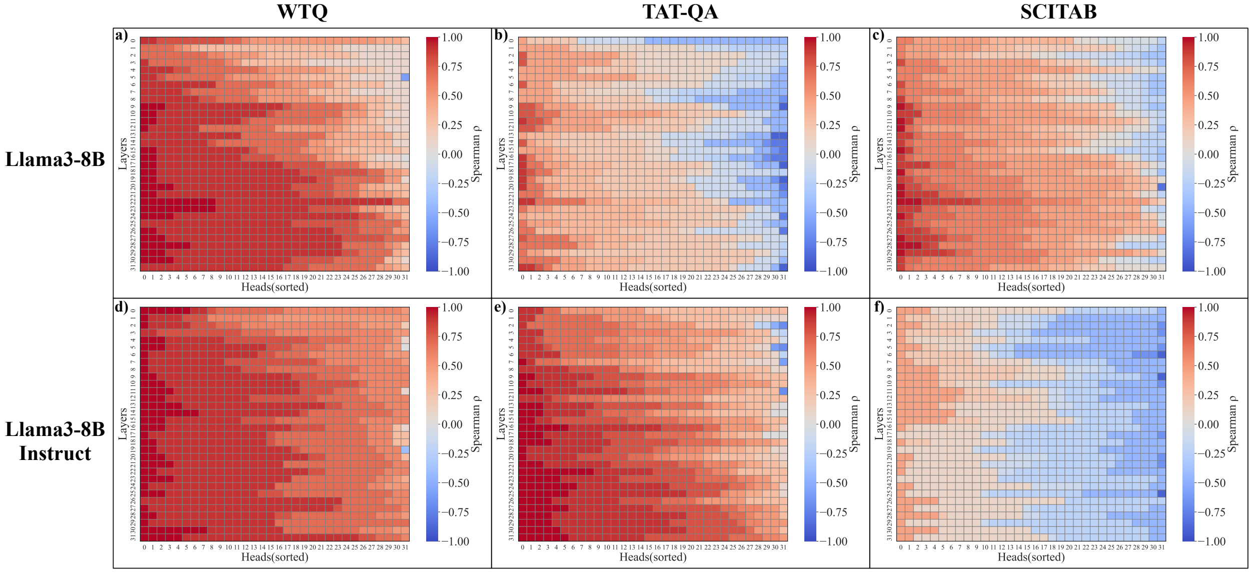


Figure 14: Heatmap showing Spearman correlation between changes in attention entropy and EM difference across all attention heads and layers in the `Llama3-8B` model(**a, b** and **c**) and `Llama3-8B-instruct` model(**d, e** and **f**).

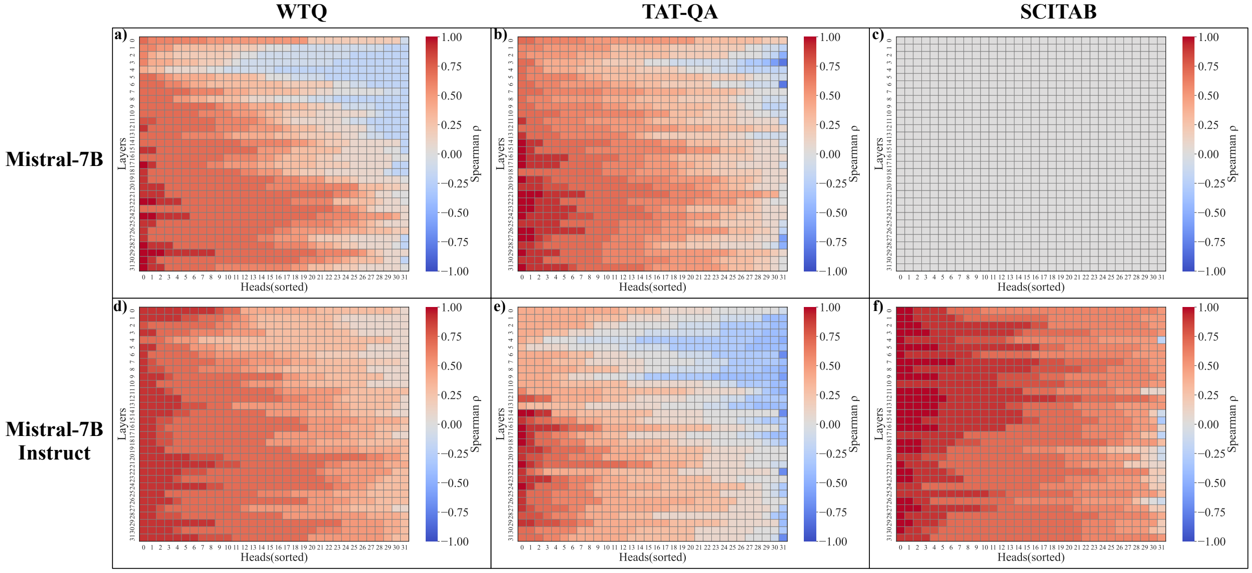


Figure 15: Heatmap showing Spearman correlation between changes in attention entropy and EM difference across all attention heads and layers in the `Mistral-7B` model(**a, b** and **c**) and `Mistral-7B-instruct` model(**d, e** and **f**).

By contrast, the chat and instruct-adapted models (`Llama3-8B-Instruct`, `Mistral-7B-Instruct`; Figure (14 and 15) subfigures **d-f**) display stronger positive correlation, with some extending beyond the middle layers. On **WTQ** in particular, both the initial encoding layers (layers $0-2$) and the highest layers (uppermost $29-31$ blocks) contribute positively to the entropy–EM relation, suggesting that conversational and instruction tuning bolsters the model's resilience to perturbations at both the token-embedding stage and the final consolidation stage. This extended positive band likely reflects enhanced parameter alignment across the architecture, enabling more stable information propagation even under input disruptions.

In general, these results yield two primary implications. First, conversational and instruction-tuning systematically extends the alignment between attention entropy instabilities and performance degradation across a broader portion of the transformer hierarchy, thereby allowing perturbation robustness not only in the mid layers but also at its boundaries. Second, task domain complexity governs which layers most critically underpin model stability: general-domain benchmarks (e.g., **WTQ**) draw on a broad spectrum of transformer depths, whereas specialized datasets remain predominantly reliant on central layers.