Table as a Modality for Large Language Models

Liyao Li*, Chao Ye*, Wentao Ye, Yifei Sun, Zhe Jiang, Haobo Wang
Jiaming Tian, Yiming Zhang, Ningtao Wang, Xing Fu, Gang Chen, Junbo Zhao
Hangzhou, China

The Challenge in Tabular LLMs

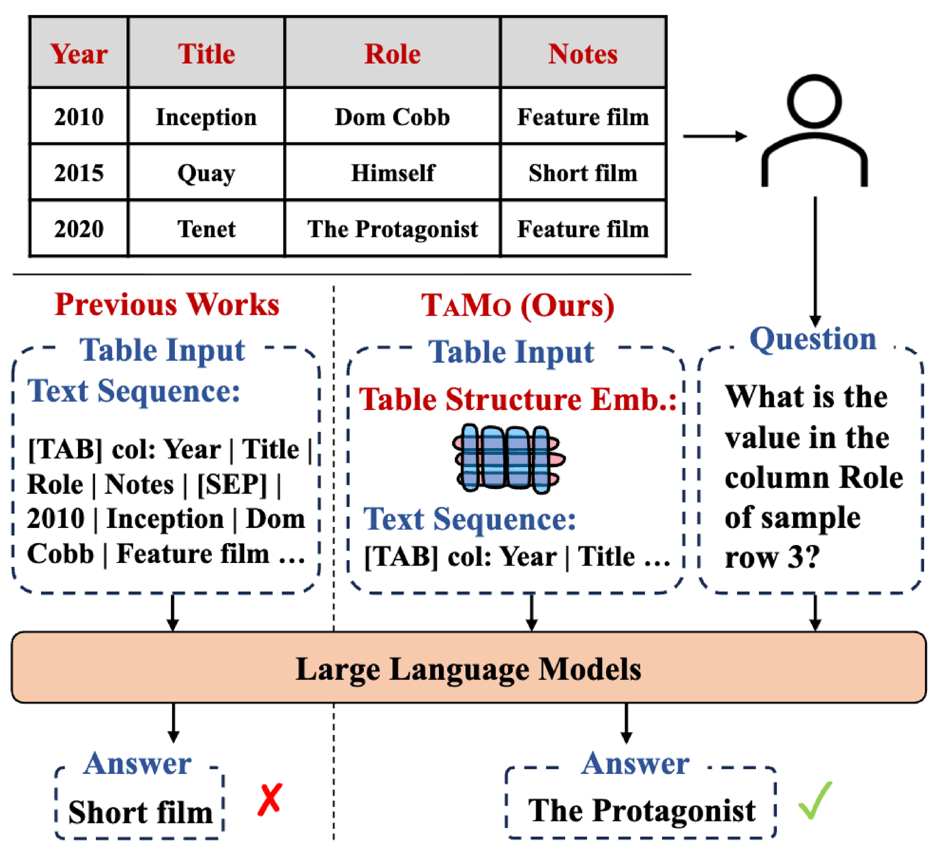
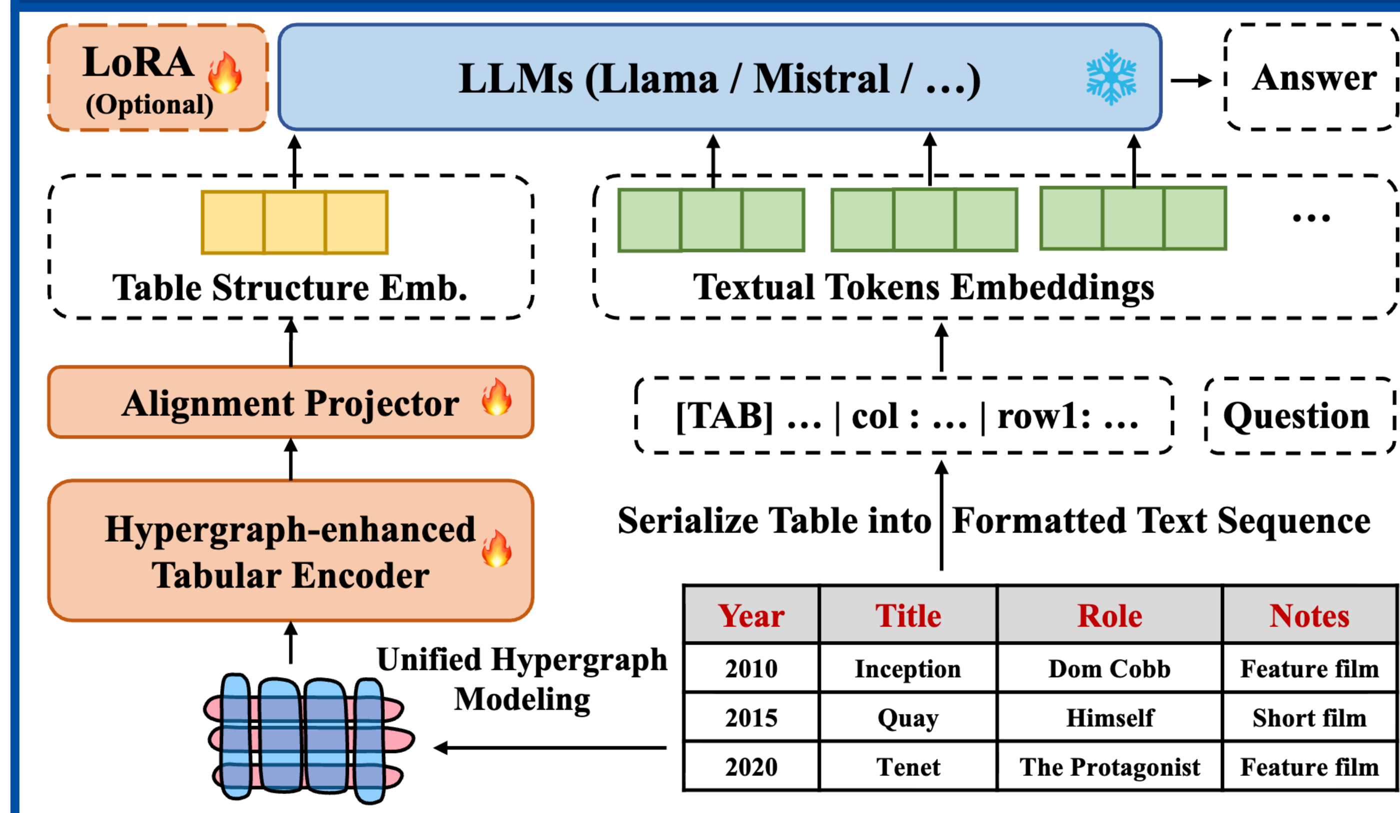


Figure 1: Current tabular LLMs oversimplifies tables into text sequences, ignoring structured information and hindering basic table cell localization tasks. This work is the first to direct table structure integration into LLMs.

- **The Serialization Bottleneck:** Current LLMs (e.g., GPT-4, Llama) treat tables as serialized 1D text sequences.
- **Loss of Structure:** This "linearization" destroys the inherent 2D structure of tables (rows, columns, hierarchies).
- **The Consequence:** Models fail to capture **Permutation Invariance**—*swapping rows or columns shouldn't change the table's meaning, but it confuses text-based LLMs.*

TAMO: A Hypergraph-Enhanced Multimodal Framework



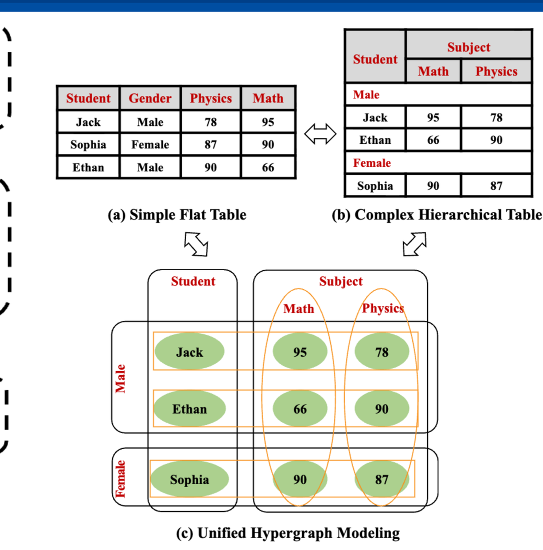


Figure 3: An example of converting arbitrary simple or complex tables into hypergraphs. A simple flat table is a special case of the complex hierarchical table. A hyperedge (e.g., table headers) in the hypergraph is a set of regular nodes. We construct the corresponding hypergraph format according to the hierarchical relationships of the table.

The TAMO framework treats tables as a separate modality, seamlessly integrated with the LLM backbone.

1. **Hypergraph-Enhanced Table Encoder:** Unlike simple embedding, we use a hypergraph structure where cells are nodes and rows/columns are hyperedges. This naturally enforces **permutation invariance** (the graph structure remains the same regardless of input order).
2. **Alignment Projector:** A learnable MLP layer maps the table structure embeddings ($\mathbf{X}_{st}$) into the LLM's semantic space.
3. **LLM Integration:** The model receives **two inputs**: the serialized textual tokens ($\mathbf{X}_{tt}$) for semantic content, and the table structure embeddings ($\mathbf{X}_{st}$) for global structural context.

Encoding & Alignment

1. Problem Formulation:
We model tables as **Hypergraphs** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to capture structural invariance1:
- **Nodes ($\mathcal{V}$):** Individual table cells (Leaf cells).
- **Hyperedges ($\mathcal{E}$):** Rows, columns, or headers containing subsets of corresponding nodes.

2. Structure Learning (Iterative Updates):
We apply permutation-invariant multiset functions to propagate information:
- **Node-to-Edge:** $\mathbf{x}_e^{t+1} = \mathrm{Fusion}\left(\mathbf{x}_e^t, \mathrm{Multiset}\left(\{\mathbf{x}_v^t | v \in e\}\right)\right)$
- **Edge-to-Node:** $\mathbf{x}_v^{t+1} = \mathrm{Multiset}\left(\{\mathbf{x}_e^{t+1} | v \in e\}\right)$

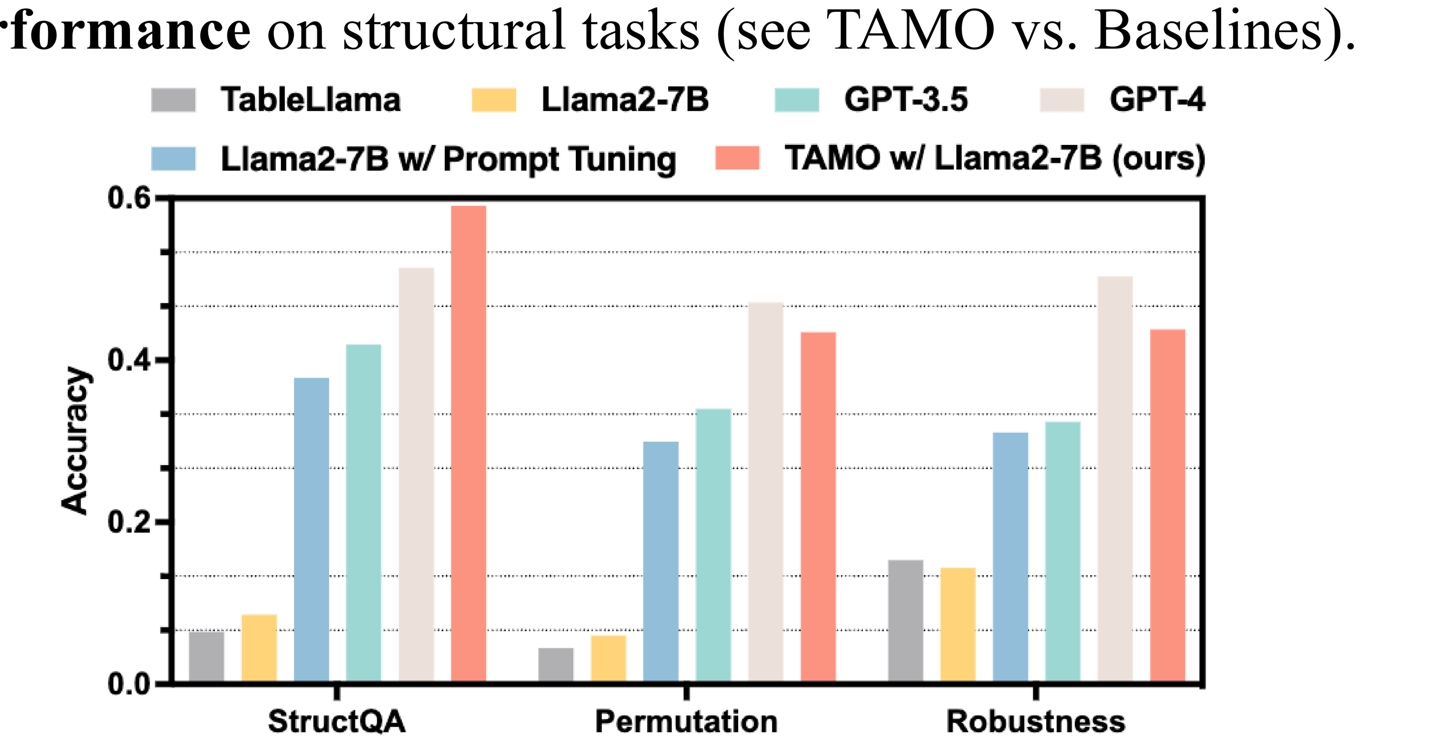
3. Alignment:
We project the learned structure ($\mathbf{X}_{st}$) to align with the serialized textual tokens ($\mathbf{X}_{tt}$) for joint reasoning:

$$\mathbf{X}_{st} = \mathrm{MLP}\left(\mathrm{Pooling}\left(\widehat{\mathbf{X}}_{\mathcal{V}}, \widehat{\mathbf{X}}_{\mathcal{E}}\right)\right)$$

$$p(\mathcal{A}|\mathcal{T}) = \prod p(a_i | \mathbf{X}_{st}, \mathbf{X}_{tt}, a_{<i})$$

Motivation and Contribution

- **Diagnostic Benchmark:** We propose **StructQA**, a benchmark focusing on tabular structure understanding. As shown in the chart, leading LLMs **fail to maintain robustness** when table rows/columns are permuted.
- **Core Philosophy:** We argue for treating tables as a unique modality—akin to visual encoders—by injecting **rich, permutation-invariant structural embeddings** to transcend the inherent limitations of text serialization.
- **The Solution:** We introduce **TAMO** (*Table as a Modality*), a framework that integrates a specialized table encoder with LLMs to preserve structural integrity, **significantly recovering performance** on structural tasks (see TAMO vs. Baselines).



Experiments with Existing Methods

- **State-of-the-Art Performance**: Evaluated on 5 benchmarks (StructQA, HiTab, WikiTQ, WikiSQL, FeTaQA). TAMO achieves an average relative gain of **42.65%**.
- **TAMO vs. Text-Only**: Significantly outperforms pure text baselines (e.g., +86.06% on HiTab).
- **TAMO vs. GPT-4**: Surpasses GPT-4 on structure-heavy tasks like StructQA and HiTab.
- **TAMO vs. Specialist SOTA**: Achieves competitive results without needing task-specific engineered architectures.

Setting	Dataset Task Type Evaluation Metric	StructQA Structural QA Accuracy	HiTab Hierarchical QA Accuracy	WikiTQ Table QA Accuracy	WikiSQL Table QA Accuracy	FetaQA Free-form QA BLEU
Inference Only	Zero-shot	8.60	7.77	14.50	21.44	20.08
Frozen LLM	Prompt tuning	37.80	26.26	29.86	61.24	29.94
	TAMO	59.07	48.86	37.06	76.45	36.52
	$\triangle_{Prompt\ tuning}$	↑ 56.27%	↑ 86.06%	↑ 24.11%	↑ 24.84%	↑ 21.98%
Tuned LLM (LoRA)	LoRA	45.67	50.76	37.13	57.10	35.80
	TAMO$^{+}_{LoRA}$	<u>70.80</u>	59.22	<u>43.53</u>	<u>84.43</u>	37.43
	$\triangle_{LoRA}$	↑ 55.03%	↑ 16.67%	↑ 17.24%	↑ 47.86%	↑ 4.55%
Tuned LLM (SFT)	TableLlama[2023b]	6.47	<u>63.76</u>	31.22	46.26	<u>38.12</u>
	SFT	62.73	54.80	43.28	79.86	37.37
	TAMO$^{+}_{SFT}$	**71.60**	**63.89**	**45.81**	**85.90**	**39.01**
	$\triangle_{SFT}$	↑ 14.14%	↑ 16.59%	↑ 5.85%	↑ 7.56%	↑ 4.39%
Others	GPT-3.5	41.93	43.62*	53.13*	41.91*	26.49*
	GPT-4	51.40	48.40*	68.40*	47.60*	21.70*
	GPT-4.1	60.33	60.54	68.14	71.21	36.75
	DeepSeek-R1	57.47	63.89	75.76	71.91	13.10
	Specialist SOTA	-	64.71[2023b]	69.10[2024]	92.07[2022]	40.50[2024]

Table 2: Results on our table structure understanding dataset *StructQA* and four table reasoning benchmarks. TAMO adds additional table modality information compared to the pure text baseline. Specialist SOTA refers to methods that design models and training tasks specifically for each dataset. "*" indicates data sourced from Zhang et al. [2023b]. The first best result for each task is highlighted in **bold** and the second best result is highlighted with an underline.

Robustness and Interpretability

- **Robustness to Permutation:**
  - In real-world tests (shuffling rows/cols), text-only LLMs show a massive performance drop.
  - **TAMO** maintains high accuracy and consistency, proving it truly understands the table structure rather than memorizing position.
- **Interpretability (Attention Map):**
  - Visualizing attention weights reveals that TAMO focuses on the **correct answer cells** ("Canada") and relevant context, whereas text-only models often get distracted by irrelevant tokens.

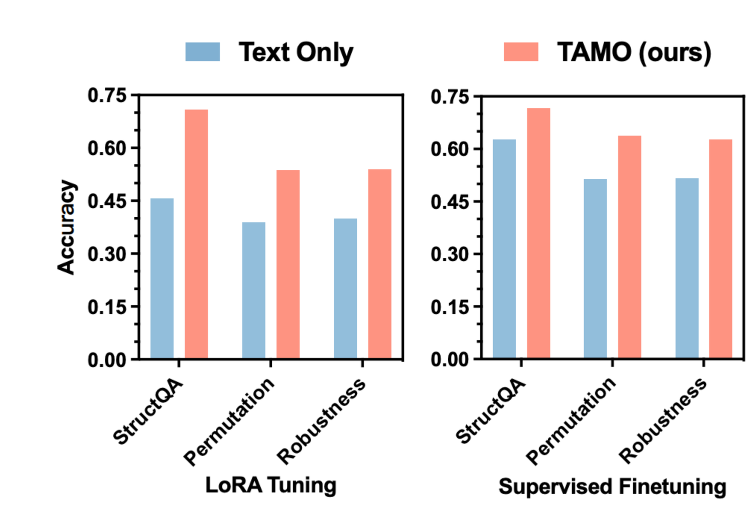


Figure 6: Evaluate the robustness of TAMO to permutation invariance on the StructQA dataset. *Permutation*: randomly permuting rows and columns in the StructQA test set. *Robustness*: the proportion of samples that remain consistent after random permutation.