Texts or Images? A Fine-grained Analysis on the Effectiveness of Input Representations and Models for Table Question Answering

Wei Zhou[1,3] **Mohsen Mesgar**[1] **Heike Adel**[2] **Annemarie Friedrich**[3]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Hochschule der Medien, Stuttgart, Germany [3]University of Augsburg, Germany
{wei.zhou|mohsen.mesgar}@de.bosch.com
annemarie.friedrich@uni-a.de adel-vu@hdm-stuttgart.de

Abstract

In table question answering (TQA), tables are encoded as either texts or images. Prior work suggests that passing images of tables to multi-modal large language models (MLLMs) performs comparably to or even better than using textual input with large language models (LLMs). However, the lack of controlled setups limits fine-grained distinctions between these approaches. In this paper, we conduct the first controlled study on the effectiveness of several combinations of table representations and models from two perspectives: question complexity and table size. We build a new benchmark based on existing TQA datasets. In a systematic analysis of seven pairs of MLLMs and LLMs, we find that the best combination of table representation and model varies across setups. We propose FRES, a method selecting table representations dynamically, and observe a 10% average performance improvement compared to using both representations indiscriminately.

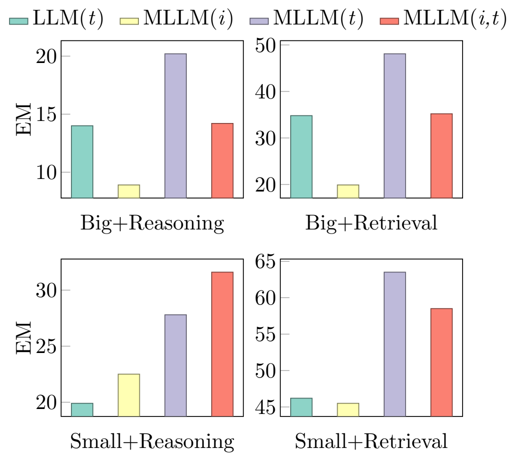


Figure 1: **Varying exact match (EM) for models and table representations under different settings** (i and t stand for image and text representations of tables). We categorize our investigation into four settings based on table size (small or big) and question complexity (retrieval or reasoning).

1 Introduction

Table is a common data format in many downstream applications (Pasupat and Liang, 2015; Parikh et al., 2020) and domains (Chen et al., 2021; Lu et al., 2023). To process table data, current approaches either pass serialized table texts to large language models (LLMs) (Herzig et al., 2020; Jiang et al., 2022; Liu et al., 2022), or input table images to multi-modal large language models (MLLMs) (Zheng et al., 2024; Alonso et al., 2024).

Deng et al. (2024) show that these two approaches yield comparable performance on table-related tasks, such as table question answering (TQA), when evaluated using GPT-4 (Achiam et al., 2023) and Gemini$_{\text{Pro}}$ (Team et al., 2023). Sometimes, passing table images to MLLMs even outperforms inputting table texts to LLMs. However, the absence of controlled setups in prior work, e.g., the investigation of performance in relation to table size, hinders a deeper understanding of their strengths and weaknesses. Moreover, existing investigations focus exclusively on closed-source models, raising the question of whether similar observations can be made using open-weights models of varying sizes.

To study the effectiveness of different combinations of table representations (image vs. text vs. a combination of both) and models (LLMs vs. MLLMs) under controlled circumstances, we build a new benchmark with four controlled settings sourced from six widely used TQA datasets. Our evaluation benchmark consists of 1600 instances, with tables in both image and text formats. We annotate instances with question complexity (retrieval vs. reasoning) and table size (small vs. big). In our experimental study, we carefully evaluate seven open-weights MLLMs with their corresponding pre-trained LLM decoders in a zero-shot way. Our results indicate that model size

Findings of the Association for Computational Linguistics: ACL 2025, pages 2307–2318
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Settings	Source Dataset (#Instance)	#Total
Retrieval		
Small	WTQ (100), TabFact (100), HiTab (200)	400
Big	WTQ (100), TabFact (100), HiTab (200)	400
Reasoning		
Small	WTQ (50), TabFact (50), HiTab (200) TabMWP (50), CRT (50)	400
Big	WTQ (50),TabFact (50), HiTab (200) TempTabTQA (50),CRT (50)	400

Table 1: Our controlled settings are based on two comparison dimensions: question complexity (retrieval vs. reasoning) and table size (small vs big). The middle column shows the source datasets and the number of their instances in our benchmark.

indeed plays a role in determining the most effective model-representation combinations: using the large Qwen-2-72B model, we find that passing table images consistently outperforms passing table texts, aligning with observations reported by Deng et al. (2024). However, for small models (with ≤12B parameters), the optimal combination of model and table representation varies across different settings, as illustrated in Figure 1. For instance, MLLMs operating on both table text and image is the most effective approach when tables are small and questions require reasoning to be solved. Key findings are presented at the end of Section 3.2. Finally, we propose **f**eature-based table **re**presentation **s**election, **FRES**, a method determining the best table representation for small MLLMs based on question complexity and table size. By applying FRES, we observe an average of 10% exact match gain compared to baseline approaches. Code and data are available.[1]

2 Benchmark Creation

We build a new evaluation benchmark consisting of 1600 instances from six common TQA datasets. The benchmark features two dimensions and provides four controlled settings, with each setting comprising 400 instances. Table 1 lists the number of instances collected for each investigated setting.

2.1 Source Datasets

We build up our benchmark from the development sets of six commonly-used single-table TQA datasets. We use table images and texts already provided in MMTab (Zheng et al., 2024) for four out of six datasets: WTQ (Pasupat and Liang, 2015), TabFact (Chen et al., 2020), TabMWP (Lu et al., 2022), and HiTab (Cheng et al., 2022). These datasets feature numerical reasoning questions. To broaden the range of reasoning types, we also include TempTabTQA (Gupta et al., 2023) for temporal reasoning and the test-only dataset CRT (Zhang et al., 2023) for commonsense reasoning. Note that TabMWP exclusively features small tables; we use it to collect instances with small tables. Similarly, TempTabTQA predominantly features large tables. We therefore use it to select instances with big tables. In TabFact, each instance comprises a table and a statement about the table. The goal is to answer whether the statement is True or False. We convert it into TQA format by decomposing statements into question-answer pairs (A.1). We generate table images for TempTabTQA and CRT from table texts, using four templates and various color encodings with HTML rendering tools (A.2). Dataset statistics are provided in A.3.

2.2 Dimensions

We categorize the instances in our benchmark by **question complexity** and **table size**. These two dimensions offer orthogonal perspectives, i.e., from a task-instruction level and from an input-data level.[2] In terms of question complexity, questions that involve identifying a verbatim answer in a table are referred to as **retrieval questions**, while those requiring additional inferences, such as numerical, temporal, or commonsense reasoning, are called **reasoning questions**. In terms of table size, we distinguish between **big** and **small** tables, motivated by the fact that MLLMs have been found to struggle with high-resolution images in visual benchmarks (Li et al., 2023).

2.3 Data Collection Process

Our goal is to collect the same number of instances for each setting from the six source datasets. In total, we collect 1600 instances, with 400 for each setting. Statistics for each setting are shown in A.3. First, we **distinguish between retrieval and reasoning questions**. Three out of six datasets (TabMWP, TempTabTQA, and CRT) exclusively feature reasoning questions. For WTQ, we differentiate retrieval questions from reasoning questions using the classification model from Zhou et al.

[1] https://github.com/boschresearch/FRES

[2] Although table structure (flat vs. hierarchical) could also be an interesting dimension, we do not condition for it because it correlates with table sizes: in our initial exploration, we find hierarchical tables are generally bigger than flat tables in existing table datasets.

MLLM	LLM	Size
Qwen-2-VL (Wang et al., 2024)	Qwen-2 (Yang et al., 2024)	7B / 72B
Pixtral (Agrawal et al., 2024)	Mistral-nemo[3]	12B
Phi-3.5-vision-instruct (Abdin et al., 2024)	Phi-3.5-mini	4B
LlaVa-Next (Li et al., 2024)	Mistral (Jiang et al., 2023)	7B
GLM-4v (Zeng et al., 2024)	GLM-4	9B
InternVL2[4]	Internlm2_5-7B-chat (Cai et al., 2024)	8B

Table 2: Selected MLLMs and LLMs. We omit citations of LLMs if they are the same as corresponding MLLMs'.

(2024) and use Qwen-2 72B (Yang et al., 2024) as the question classifier, which achieves an accuracy of 93% on a set of annotated instances from Zhou et al. (2024). Details are available in A.4. For TabFact and HiTab, we leverage the datasets' annotations on question types.

Next, we **distinguish between big and small tables.** Thresholds determining size groups are calculated as the averages of number of pixels and number of table text tokens in MMTAB (Zheng et al., 2024). We select MMTAB for threshold calculation as it features a vast collection of more than 120k tables originally from web pages and reports, targeting human readers. Size distribution of different datasets in MMTab can be found in A.5. Tables with both pixel numbers and token numbers smaller than the average (2e6 and 288) are classified as small. Tables with both values larger than averages are categorized as big. The rest of the tables ($\sim$22%) are not included in the benchmark. This ensures each size group features big/small tables in both image and text representations.

After distinguishing question complexity and table size, we rank instances with the same question type by their table sizes, and choose the top k (shown in Table 1) instances from each source dataset in ascending order to collect data with small tables and descending order for big tables. This way, we ensure that the distribution of data featuring big/small tables is distinctive. We also keep a balanced distribution of different table structures (800 instances with flat/hierarchical tables).

[3]https://mistral.ai/news/mistral-nemo/

[4]https://internvl.github.io/blog/2024-07-02-InternVL-2.0/

3 Experiments

We aim to compare different combinations of table representations and model types. In this section, we present selected models and evaluation metrics, as well as discuss the results.

3.1 Models and Evaluation

We select **seven open-weight** MLLMs and their corresponding pre-trained LLM decoders (shown in Table 2) for reproducibility and generalizability. The selected MLLMs contain a significantly smaller visual encoder compared to the LLM decoder. This ensures that MLLMs and LLMs share the same model structure with only minimal differences in parameters.

To mitigate pre-training data contamination, i.e., ensure that the curated evaluation data has not been greatly exposed to a model during pre-training, we only select a pair of MLLM and LLM if both feature a low accuracy ($\leq$ 20%) in all four controlled settings when masking out questions, and when masking out tables. Model performances with questions or tables masked can be found in A.6.

In total, we select **six** pairs of **small** models (MLLMs and LLMs), and **one** pair of **big** models (due to a limited number of large MLLMs satisfying the above model selection conditions). We use the same table layouts and prompt templates (A.7) as previous work (Deng et al., 2024; Zheng et al., 2024). During **evaluation**, we further exclude all instances that any of the models predict correctly when only seeing the table or question. In total, 84% of the evaluation data is preserved (at least 80% per setting), ensuring a sufficient amount of data available for evaluation. We use exact match accuracy (EM) between reference answers and predicted answers as the evaluation metric and the Wilcoxon signed-rank test (Wilcoxon, 1945) for testing statistical significance.

3.2 Results

We present the averaged results across all six small models in Figure 1. Results for individual models are shown in A.8. Table 3 shows the results for Qwen-2-72B and Qwen-2-VL-72B.

Previous work suggests that passing table images to MLLMs can result in comparable or even better performances than inputting table texts to LLMs (Deng et al., 2024). We find that the observation is valid only with large models: in Table 3, we observe that MLLM(i) consistently leads to better

Model	B/Reason	B/Retrieve	S/Reason	S/Retrieve
LLM (*t*)	29.2	69.5	36.9	81.3
MLLM (*i*)	46.5	75.4	67.1	87.7
MLLM (*t*)	46.8	**78.9**	60.7	88.1
MLLM (*i*,*t*)	**51.5**	78.2	**70.0**	**89.4**

Table 3: Exact Match of Qwen-2-72B on four different settings. *i*, *t* refer to tables as images and texts, respectively. B and S stand for Big and Small in terms of table size, respectively.

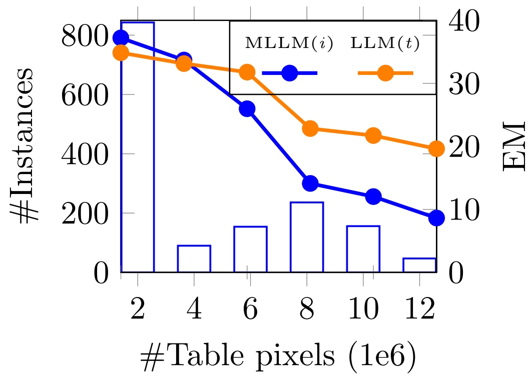


Figure 2: Evaluation of table size robustness. The bar plot shows the number of instances sampled for each bin, and the line plots show the performance of different approaches against varying table sizes.

performance than LLM(t), regardless of question complexity and table size.

However, **the performance of these two approaches differs a lot under different settings when evaluating with small-sized models**: LLMs using table text (LLM(t)) outperform MLLMs using table images (MLLM(i)) when tables are large ($p < .05$). We believe this is because MLLMs struggle with processing large images (Li et al., 2023). To prove that, we categorize instances into six distinctive bins according to their table sizes, and plot the EM of different approaches using grouped instances. This is shown in Figure 2. We find that table image representations are less robust compared to text representations with regard to table size. A similar figure plotted against table token numbers is shown in Figure 5, conveying the same message.

For small tables, question complexity matters: MLLMs excel at reasoning questions using table images, while we do not find statistically significant differences for retrieval questions ($p = .67$). This provides evidence for the hypothesis proposed in Deng et al. (2024) that representing tables as images can help LLMs in complex reasoning. We suspect MLLMs' reasoning advantages come from training on massive image reasoning data.

When comparing general capabilities of MLLMs and LLMs in solving TQA with textual inputs, i.e., LLM(t) versus MLLM(t), we find that MLLMs outperform LLMs. It appears that continuing to pre-train and fine-tune LLMs with a small vision encoder using multi-modal data further enhances models' abilities in solving TQA.

By comparing different table representations with MLLMs, we observe that text-based approaches excel with large tables ($p < .05$). However with small tables, combining both table representations is the most effective choice for reasoning questions, while text alone suffices for retrieval questions ($p < .05$). We suspect that passing both table representations better triggers the reasoning abilities of MMLMs, thus leading to superior performance on reasoning questions. However, in terms of information retrieval, the input is best represented as text, probably due to the task being mostly presented in textual formats during instruction-tuning (Zhu et al., 2023).

Key Takeaways

- For the large 72B model, table image is a more effective representation than table text. Passing both representations to the MLM results in the best performance.
- For small models:
  - with big tables, using text representations with MLLMs leads to the best performance.
  - with small tables, if the question type is reasoning, providing MLLMs with both table representations results in optimal performance. For retrieval questions, supplying table texts to MLLMs suffices.

4 FRES: Feature-Based Table Representation Selection

Small MLLMs require less memory and are faster during inference than big MLLMs. We use observations found before and propose FRES, a method selecting the best table representations when using small-sized models: big tables are passed as texts to MLLMs; small tables are processed as texts when questions are retrieval-based; otherwise, both representations are passed to MLLMs. We employ

Methods	WTQ	TabFact	HiTab	WiKiSQL
Pxl-*t*	51.3	75.7	59.7	58.8
Pxl-*i*	42.2	74.2	40.8	51.7
Pxl-*t,i*	52.5	**75.9**	62.2	60.1
Pxl-FRES	**54.4**	75.4	**64.0**	**61.4**
TL-*t*	34.4	**63.7**	44.3	47.8
TL-*i*	19.2	59.5	10.5	31.6
TL-*t,i*	17.2	60.9	16.3	29.5
TL-FRES	**38.8**	63.1	**49.7**	**48.4**

Table 4: Exact Match of models on TQA datasets. t, i refer to tables as texts and as images, respectively. Pxl and TL refer to Pixtral and TableLlaVA, respectively.

the same classifier used in Section 2.3 to determine question types. We regard tables with either pixels or token numbers larger than values in 2.3 as big. Otherwise, they are classified as small.

Datasets. We evaluate our method with the **test sets** of three aforementioned TQA benchmarks: WTQ, TabFact (small test) and HiTab for their diverse table sizes and question types. We also include one additional dataset, WikiSQL (Zhong et al., 2017), that has not been used in our analysis to test the generalizability of our findings. Dataset statistics are presented in A.3.

Baselines and Models. We test four approaches: passing tables as **texts**, as **images**, as **both** representations, and the ones decided by FRES. We choose the best performing MLLM from our previous analysis: **Pixtral 12b** (see A.8 for individual model's performance) as well as a fine-tuned MLLM for TQA: **TableLlaVA 7B** (Zheng et al., 2024).

Results. Table 4 displays the Exact Match of various approaches. Our proposed method, which selects input table formats based on question complexity and table size, proves to be effective. It achieves an average 10% of EM gain across two models and four datasets, compared to using both representations indiscriminately. In addition, it improves efficiency by reducing input token numbers (A.9). We do not observe big differences between the baselines and our method on TabFact. This might stem from the dataset's simplicity: TabFact is a classification dataset with only two labels. It is likely that differences between each representation become small or even diminish.

Error Analysis. We analyze FRES in terms of its failure patterns. To do that, we sample 200 instances where wrong predictions were given by Pixtral-7B with FRES, with each test dataset 50 error instances. To investigate the impacts of table representations in causing errors, we obtain predictions of the same instances using different representations other than the one selected by FRES. We note down changes in prediction correctness. If all representations fail to result in a correct prediction, we regard the error as coming from the **limited capability of the model** itself. As FRES do not decide for image-only representations, any correct predictions obtained by changing to image-only table representations are regarded as **exceptional cases**. If changing table representations leads to correct predictions and the question type is predicted correctly, we regard the error as coming from **limitation of table size thresholds**. Otherwise, we regard the error coming from **failures in the question classification model**. To obtain correct question type labels, we manually annotate the 200 instances. We find 65% of the errors fall into the category of limited model capability, while approximately 10% are exceptional cases in which our observations do not capture. Around 19% of the errors stem from question classification errors, and 7% from limitations of table size thresholds.

5 Conclusions

We presented a new benchmark for exploring the impact of using different input formats (text vs. image) with LLMs vs. MLLMs conditioned on question complexity and table size. In a systematic evaluation of seven models, we find that each combination of table representation and model type performs differently under different settings. We propose FRES, a method integrating our findings to dynamically select the best table representation. FRES effectively improves TQA performance.

Limitation

This study primarily focuses on table question answering, motivated by the availability of extensive datasets. However, it could be pertinent to investigate whether our findings are applicable to other table-centric tasks, such as table summarization. Moreover, we only explore the application of our findings in a simple setting where MLLMs are prompted in a zero-shot manner. Future work can develop more intricate methods leveraging both modalities with our findings. Lastly, we assume the presence of tables in both image and textual formats for our initial exploration. Yet, in practical applications, the conversion between these representations may result in the loss of information, e.g., OCR might not work well when converting an image of a big table image to corresponding textual representation (Zheng et al., 2024).

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Young Jin Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Praneetha Vaddamanu, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng-Yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Singh Chaplot, Jessica Chudnovsky, Saurabh Garg, Théophile Gervet, Soham Ghosh, Am'elie H'eliou, Paul Jacob, Albert Q. Jiang, Timothée Lacroix, Guillaume Lample, Diego de Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Reddy Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, and Thomas Wang. 2024. Pixtral 12b. *ArXiv*, abs/2410.07073.

Iñigo Alonso, Eneko Agirre, and Mirella Lapata. 2024. PixT3: Pixel-based table-to-text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6721–6736, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiao wen Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhen Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kui-Jie Liu, Xiaoran Liu, Chen Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xing Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Rui Ze Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fen-Fang Zhou, Zaida Zhou, Jingming Zhuo, Yi-Ling Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *ArXiv*, abs/2403.17297.

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

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

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland. Association for Computational Linguistics.

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

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453, Singapore. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

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

Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. OmniTab: Pretraining with natural and synthetic data for few-shot table-based question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 932–942, Seattle, United States. Association for Computational Linguistics.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *ArXiv*, abs/2407.07895.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023. Monkey: Image resolution and text label are important things for large multi-modal models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26753–26763.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. TAPEX: Table pre-training via learning a neural SQL executor. In *International Conference on Learning Representations*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and A. Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *ArXiv*, abs/2209.14610.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv*, abs/2409.12191.

Frank. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, 1:196–202.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. 2024. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Team Glm Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Ming yue Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiaoyu Xia, Xiaohan

Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yi An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhenyi Yang, Zhengxiao Du, Zhen-Ping Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *ArXiv*, abs/2406.12793.

Zhehao Zhang, Xitao Li, Yan Gao, and Jian-Guang Lou. 2023. CRT-QA: A dataset of complex reasoning question answering over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2131–2153, Singapore. Association for Computational Linguistics.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

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

Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. 2024. FREB-TQA: A fine-grained robustness evaluation benchmark for table question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2479–2497, Mexico City, Mexico. Association for Computational Linguistics.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. 2023. Large language models for information retrieval: A survey. *ArXiv*, abs/2308.07107.

A Appendix

A.1 Converting TabFact to a TQA Setting

Each instance in TabFact comprises a table, a statement about the table, and an answer if the statement is True or False. We select only the true statements to ensure that the answers contained in the statements are correct. Subsequently, we use GPT-4 to decompose these statements into question-answer pairs. We exclude instances whose answers are not found in the original statements. To test the validity of the created question answering dataset, we run GPT-4 three times and choose the most frequent answers as the predicted answers. By comparing the predicted answers and the gold answers obtained from GPT-4 by statement decomposing, an accuracy of 90% is reached. In cases of errors, we find that either the evaluation is challenging (due to free-form text) or the original statements are inaccurate. This suggests the validity of the question answering dataset created from TabFact.

A.2 Creating Table Image

We use four templates to create table images given table text. The templates are shown in Figure 3. To create a table image, we first represent tables as Pandas DataFrame object, then we parse a DataFrame into HTML with the function `df.to_html`. We convert HTML to image with Python package `Html2Image`.[5] Lastly, we crop images to eliminate redundant white space.

A.3 Dataset Collection and Statistics

Evaluation Data Statistics. To create our evaluation data, we have to distinguish retrieval and reasoning questions, as well as small and big tables. We present an analysis of our collected evaluation data in two aspects: **statistics per source dataset**, and **statistics per evaluation setting**. These are shown in Table 5 and 6, respectively.

Test Data Statistics. We present statistics for four test datasets in Table 8. We show averages of cell number, table tokens, and resolution, together with the percentage of small tables, reasoning questions, and instances featuring both small tables and reasoning questions. As no annotation of question complexity exists, the statistics are calculated based on the question classifier explained in A.4. The categorization of small and large tables is based on heuristics calculated from MMTab (A.5).

[5] https://pypi.org/project/html2image/

Dataset	Q Length	T Length	A Length	T Resolution	Q Complexity	T Structure	#Samples
TABMWP	25.1	21.7	1	2.4e4	numerical reasoning	flat	50
TempTabTQA	13.4	461.9	1.7	3.8e6	temporal reasoning	flat	50
CRT	24.2	304.8	1.3	4.6e6	mixed reasoning	flat	100
WTQ	10.2	676	1.9	3e6	retrieval, reasoning	flat	300
TabFact	11.3	320	1.8	3.4e6	retrieval, reasoning	flat	300
HiTab	17.6	425.8	1.6	4.7e6	retrieval, reasoning	hierarchical	800
All	15.6	433.9	1.7	4e6	retrieval, reasoning	flat, hierarchical	1600

Table 5: General dataset features and statistics. Q, T, and A stands for question, table, and answer, respectively. Lengths and resolutions are averaged over collect samples. Q, T, and A lengths are calculated as the number of white-space separated tokens.

Setting	Q Length	T Length	A Length	T Resolution	#Samples
Retrieve-Small	13.3	105.9	1.5	6.8e5	400
Retrieve-Big	14.3	844.5	1.9	7.1e6	400
Reasoning-Small	18.0	93.0	1.5	6.8e5	400
Reasoning-Big	16.7	692.1	1.9	7.6e6	400

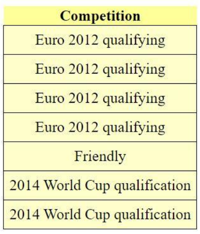
Table 6: Statistics per evaluation setting. Q, T, and A stands for question, table, and answer, respectively. Lengths and resolutions are averaged over collected samples. Q, T, and A lengths are calculated as the number of white-space separated tokens.

Year	Regular Season
2008	2nd, Mid Atlantic
2009	5th, Atlantic
2010	5th, Atlantic
2010–11	In progress
2011	4th, Atlantic
2012	3rd, Atlantic
2013	3rd, Atlantic

(a) Table without cell border/DataFrame-like.

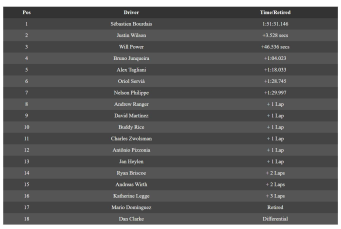
Athlete	Race 1\nTime	Race 2\nTime
Stefan Shalamanov	DNF	–
Borislav Dimitrachkov	DNF	–
Lyubomir Popov	1:10.73	DNF
Stefan Shalamanov	58.68	53.69
Lyubomir Popov	57.78	53.03
Borislav Dimitrachkov	57.58	53.23
Petar Popangelov	55.14	51.20

(b) Table with cell boarders.



Competition
Euro 2012 qualifying
Euro 2012 qualifying
Euro 2012 qualifying
Euro 2012 qualifying
Friendly
2014 World Cup qualification
2014 World Cup qualification

(c) Table with partial borders.



Pos	Driver	Time/Retired
1	Sébastien Bourdais	1:51:31.146
2	Justin Wilson	+3.528 secs
3	Will Power	+46.536 secs
4	Bruno Junqueira	+1:04.023
5	Alex Tagliani	+1:18.033
6	Oriol Servià	+1:28.745
7	Nelson Philippe	+1:29.997
8	Andrew Ranger	+ 1 Lap
9	David Martínez	+ 1 Lap
10	Buddy Rice	+ 1 Lap
11	Charles Zwolsman	+ 1 Lap
12	Antônio Pizzonia	+ 1 Lap
13	Jan Heylen	+ 1 Lap
14	Ryan Briscoe	+ 2 Laps
15	Andreas Wirth	+ 2 Laps
16	Katherine Legge	+ 3 Laps
17	Mario Domínguez	Retired
18	Dan Clarke	Differential

(d) Table with wide cell space.

Figure 3: Different table templates.

Dataset Licenses We build our evaluation dataset based on a subset of MMTab (Zheng et al., 2024), CRT (Zhang et al., 2023) and TempTabTQA (Gupta et al., 2023). The three datasets are publicly available under the licenses of APACHE-2.0[6], MIT[7] and CC-BY-4.0[8], respectively. In terms of the test datasets: WTQ (Pasupat and Liang, 2015), TabFact (Chen et al., 2020), HiTab (Cheng et al., 2022) and WikiSQL (Zhong et al., 2017), they are under the license of CC-BY-SA-4.0[9], MIT, BSD-3 CLAUSE[10] and C-UDA[11] respectively.

A.4 Question Type Classification

To distinguish retrieval and reasoning questions in **TabFact**, we use the simple and complex splits to collect instances, respectively. For **HiTab**, we utilize the dataset's annotations to differentiate between retrieval and reasoning questions; specifically, questions categorized with an aggregation type of *None* are considered retrieval questions, while all others are deemed reasoning questions. For **WTQ**, we utilize the question type classifier proposed in Zhou et al. (2024). More specifically, a rule-based method is applied first: if an answer is

[6] https://opensource.org/license/apache-2-0

[7] https://opensource.org/license/mit/

[8] https://creativecommons.org/licenses/by/4.0/legalcode

[9] https://creativecommons.org/licenses/by-sa/4.0/

[10] https://opensource.org/license/bsd-3-clause

[11] https://github.com/microsoft/HiTab?tab=License-1-ov-file

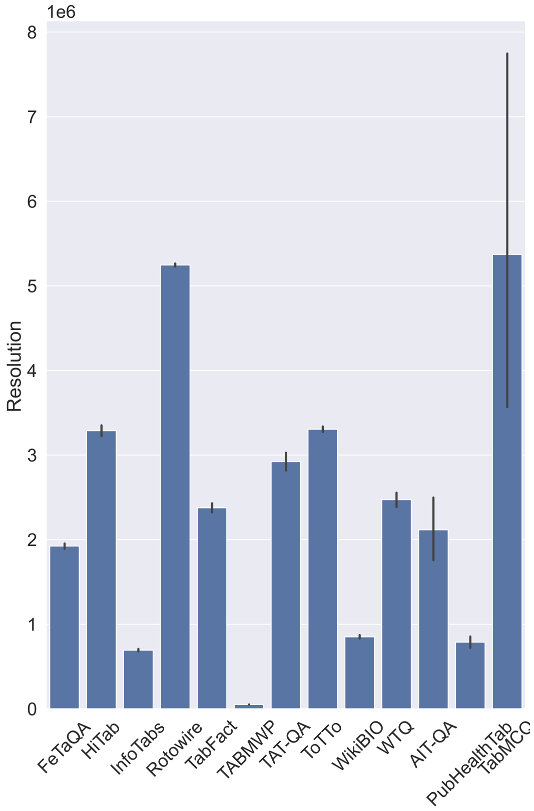


Figure 4: Resolution distribution of MMTab.

not in a table, a question is classified as a reasoning question. If a question contains comparative terms (we detect it using NLTK), the question is classified as a reasoning question. Next, an LLM takes in a question and a table and returns a predicted question type. We replace the LlaMA 2 (13b) used in the original paper with Qwen 2(72b) for its better general capabilities, but keep the prompt the same. We examine the validity of the question classifier by testing it on 200 instances from WTQ, annotated with gold question type (Zhou et al., 2024). This results in an accuracy of 93%.

A.5 Table Size threshold

We use the entire MMTab dataset (Zheng et al., 2024), consisting of more than 120 thousand table images to calculate the averaged table image resolution. This results in a value of 2e6 (width*height). The resolution distribution is shown in Figure 4. We average all token numbers in the table texts presented in MMTab. This results in an average of 288 tokens and 120 cells.

A.6 Model Selection

Table 7 shows the number of instances that models still predict correctly when masking out questions or tables under different settings. We eliminate models when either MLLMs or corresponding LLMs can predict 20% of the data correctly in any setting. This means we do not consider models that can predict more than 400*0.2=80 instances. All six examined models passed the test.

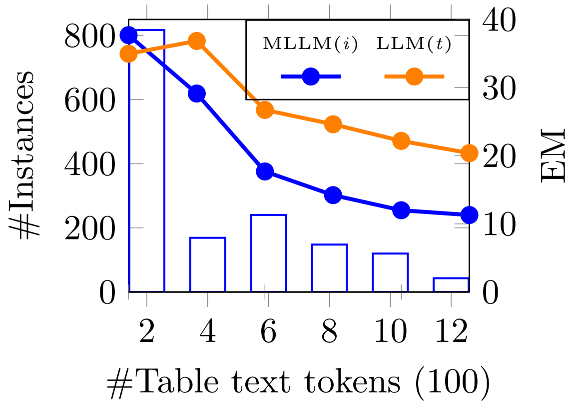


Figure 5: Evaluation of table size robustness. The bar plot shows the number of instances sampled for each bin and the line plots show the performance of different approaches against varying table sizes.

A.7 Prompts

We present prompts and table formats used in this study in Figure 6.

A.8 MLLM Result

Table 9 shows the performances of individual MLLMs on the evaluation set. We find that Pixtral 12B performs the best among all evaluated small models.

A.9 Efficiency Analysis of FRES

We compare FRES with passing both table and image representations to MLLMs indiscriminately in terms of efficiency. FRES avoids passing table images for around 80% of data in the four test sets. This results in up to 66% fewer input token numbers over the six investigated small-size models, compared to passing both representations.

Models	No Question				No Table			
	Big/Reasoning	Big/Retrieve	Small/Reasoning	Small/Retrieve	Big/Reasoning	Big/Retrieve	Small/Reasoning	Small/Retrieve
Mistral-7b	9	6	16	8	16	34	4	18
LlaVA-Next	0	0	3	0	21	49	4	24
Mistral-Nemo-12b	2	1	9	2	19	62	6	27
Pixtral-12b	0	0	1	0	37	57	3	24
Qwen2-7b	10	9	32	2	25	53	4	19
Qwen2-VL-7b	2	2	3	0	29	61	4	30
Phi-3.5-mini	4	0	4	2	4	11	0	6
Phi-3.5-vision	1	2	1	0	14	37	2	14
GLM-4-9b	0	1	7	1	3	31	0	10
GLM-4v-9b	2	3	13	1	28	28	8	26
Internlm2_5-7b	7	4	14	3	10	34	0	8
InternVL2-8b	1	2	3	0	24	66	3	26

Table 7: Number of instances that models still correctly predict when making out questions or tables.

Prompt 1: Please inspect the table(s) and then provide an answer to the question. Besides, your final answer should be in the JSON format {"answer": [<a list of answer strings>]} such as {"answer": ["87.56", "12.43"]}. Directly return the final answer and nothing else.
Table: {table}
Question:{question}

Prompt 2: Let's pretend you are an expert in reading table and answer questions. Return the answer in JSON format {"answer": [<a list of answer strings>]} such as {"answer": ["87.56", "12.43"]}. Directly return the final answer and nothing else.
Table: {table}
Question:{question}

Prompt 3: Please think step by step and return the answer in JSON format {"answer": [<a list of answer strings>]} such as {"answer": ["87.56", "12.43"]}.
Table: {table}
Question:{question}

```
Team | Rank
  A  |  1
  B  |  2
```

Table format 1

[“Team”, “Rank”], [“A”, “ 1”], [“B” , “2”]

Table format 2

```
Team, Rank
[Row 1]:  A,  1
[Row 2]:  B,  2
```

Table format 3

Figure 6: Prompts and table formats.

Dataset	#cell	#table-text	#table-img	%small table	%reasoning question	%small_reasoning
WTQ	165	429	2.4e6	57.7	71.4	41.6
TabFact	94	258	1.8e6	69.3	51.2	35.8
HiTab	180	360	3.5e6	54.5	46.0	26.0
WikiSQL	95	227	1.7e6	71.3	29.8	22.0

Table 8: Statistics of test datasets. We show the average number of cells, white-space-separated table tokens, and image resolutions (height*width). In addition. Percentages of small tables, reasoning questions as well as a combination of both are also presented.

Model	Setting	Big/Reasoning	Big/Retrieval	Small/Reasoning	Small/Retrieval
Qwen2 7b	LLM (*t*)	15.6	41.3	20.6	59.7
	MLLM (*i*)	13.1	25.2	26.8	72.2
	MLLM (*t*)	21.6	54.5	35.3	73.6
	MLLM (*i*,*t*)	9.1	20.7	36.0	72.0
Pixtral 12b	LLM (*t*)	19.4	58.7	29.0	73.3
	MLLM (*i*)	17.3	36.9	41.7	71.6
	MLLM (*t*)	**30.3**	63.1	46.1	**75.6**
	MLLM (*i*,*t*)	29.8	**65.7**	**48.5**	73.5
Phi-3.5 4b	LLM (*t*)	12.9	39.8	20.9	55.5
	MLLM (*i*)	11.2	35.6	19.4	58.7
	MLLM *t*)	18.3	46.2	21.4	68.3
	MLLM (*i*,*t*)	16.6	40.3	36.1	67.2
LlaVA 7b	LLM (*t*)	12.9	33.1	11.7	38.7
	MLLM (*i*)	1.3	3.3	4.2	14.7
	MLLM (*t*)	14.6	44.0	15.4	59.0
	MLLM (*i*,*t*)	7.4	24.0	16.6	45.6
GLM-4 9b	LLM (*t*)	11.6	18.3	12.7	24.3
	MLLM (*i*)	5.9	12.0	14.9	27.7
	MLLM (*t*)	16.7	25.8	13.7	36.9
	MLLM (*i*,*t*)	2.9	9.3	15.5	27.4
Intern 8b	LLM (*t*)	11.7	17.4	24.6	25.7
	MLLM (*i*)	4.7	6.6	28.0	43.1
	MLLM (*t*)	19.8	55.0	35.1	67.7
	MLLM (*i*,*t*)	19.1	51.4	36.8	65.2
Average	LLM (*t*)	14.0	34.8	19.9	46.2
	MLLM (*i*)	8.9	19.9	22.5	45.5
	MLLM (*t*)	20.2	48.1	27.8	63.5
	MLLM (*i*,*t*)	14.2	35.2	31.6	58.5

Table 9: Exact Match of different models on four different settings. *i*, *t* refer to tables as image and text representations.