

Does Table Source Matter? Benchmarking and Improving Multimodal Scientific Table Understanding and Reasoning

Anonymous ACL submission

Abstract

Recent large language models (LLMs) have advanced table understanding capabilities but rely on converting tables into text sequences. While multimodal large language models (MLLMs) enable direct visual processing, they face limitations in handling scientific tables due to fixed input image resolutions and insufficient numerical reasoning capabilities. To address these challenges, we present MMSci, a comprehensive dataset for scientific table understanding and reasoning. MMSci consists of three key components: (1) MMSci-Pre, a domain-specific dataset of 52K scientific table structure recognition samples, (2) MMSci-Ins, an instruction tuning dataset with 12K samples across three table-based tasks, and (3) MMSci-Eval, a benchmark with 3,114 testing samples specifically designed to evaluate numerical reasoning capabilities. Based on MMSci, we leverage and evaluate MLLMs with dynamic input resolution capabilities for scientific table understanding. Extensive experiments demonstrate that our domain-specific approach with 52K scientific table images achieves superior performance compared to 150K general-domain tables, highlighting the importance of data quality over quantity. Our proposed framework shows significant improvements in both general table understanding and numerical reasoning capabilities, with strong generalisation to held-out datasets. Our code and data are publicly available at https://anonymous.4open.science/r/MMSci_Table-F278/.

1 Introduction

Tables serve as a fundamental tool for organising structured information across diverse domains. Recent studies have shown the potential of leveraging large language models (LLMs) to automatically understand and process tabular data, which has emerged as a critical research direction with applications such as Table Question Answering (TQA) (Pasupat and Liang, 2015), Table Fact Veri-

fication (TFV) (Chen et al., 2020a), and Table-to-Text Generation (T2T) (Moosavi et al., 2021).

However, current table-oriented LLMs (Zhang et al., 2023; Li et al., 2023b) face inherent limitations as they require converting tables into sequential text formats (i.e., HTML strings), potentially losing crucial structural and positional information. While table-based multimodal large language models (MLLMs) have addressed this by enabling direct processing of table images, several critical limitations persist: (1) fixed input image resolutions that constrain practical applicability, (2) limited capability in processing scientific tables that contain significant numerical values, and (3) insufficient numerical reasoning abilities for scientific domain tasks. These limitations are particularly significant in scientific domains, where tables frequently incorporate complex numerical relationships.

Recent MLLMs have demonstrated success with general tables but struggle with scientific tables due to their dense numerical content and complex reasoning requirements. Our work demonstrates that domain-specific data quality significantly outperforms quantity, challenging conventional scaling laws in multimodal learning. Scientific table numerical reasoning requires multi-step operations including addition, subtraction, comparison, and other mathematical operations to derive conclusions from tabular data, going beyond simple fact extraction. Current MLLMs, however, lack the specific training data to handle these sophisticated scientific table understanding and reasoning requirements.

To address these challenges, we introduce MMSci, a comprehensive dataset for scientific table understanding and reasoning. We first conduct a systematic analysis of table source effectiveness through MMSci-Pre, a carefully curated dataset containing 52K structure recognition samples derived from scientific papers. Our experimental results demonstrate that MLLMs trained on these

domain-specific table images significantly outperform those trained on 150K general-domain tables, establishing the importance of data quality over quantity in table understanding tasks.

Building upon this foundation, we then create MMSci-Ins, an instruction tuning dataset comprising 12K samples with explicit intermediate reasoning steps across three fundamental tasks: TQA, TFV, and T2T. Each sample includes detailed step-by-step reasoning processes to develop models’ table-based numerical reasoning and scientific analysis capabilities. To overcome the limitations of fixed-resolution approaches in existing table MLLMs (Lee et al., 2023; Alonso et al., 2024; Zheng et al., 2024), we leverage and evaluate existing dynamic input resolution capabilities across different model architectures (Qwen2-VL-7B-Instruct and LLaVA-NeXT-7B). Our analysis reveals that mere technical capability for dynamic resolution is insufficient without proper cross-modal alignment. Experimental results demonstrate consistent performance improvements across both general table understanding and specialised numerical reasoning tasks.

To enable comprehensive evaluation, we establish MMSci-Eval, a benchmark with 3,114 testing samples requiring numerical reasoning capabilities. The benchmark provides rigorous assessment of models’ performance across TQA, TFV, and T2T tasks. Our extensive experiments demonstrate that our 52K scientific table images prove more effective than 150K general-domain table images for both general understanding and numerical reasoning tasks. This efficiency highlights the value of domain-specific, high-quality data in developing robust table understanding capabilities.

Our contributions are summarised as follows:

- We introduce MMSci, a comprehensive dataset consisting of three components: (1) MMSci-Pre, consists of 52K table image-to-HTML table structure recognition samples; (2) MMSci-Ins, an instruction tuning dataset of 12K samples with reasoning steps; and (3) MMSci-Eval, a benchmark with 3,114 samples for numerical reasoning capabilities assessment across TQA, TFV, and T2T tasks.
- We develop a comprehensive table-based MLLM framework that achieves strong performance on three table-based numerical reasoning tasks while demonstrating robust generalisation to held-out datasets.

- We implement dynamic input resolution capabilities across different model architectures, validating the effectiveness of our approach through consistent performance gains on both Qwen2-VL-7B-Instruct and LLaVA-NeXT-7B.

2 Related Work

2.1 Table Understanding Models

Early table-based models based on general language models with large-scale table corpus (Liu et al., 2022; Chen et al., 2023) only support limited types of tables and tasks. Table understanding capabilities have been enhanced through prompt engineering (Chen, 2023; Sui et al., 2023), instruction tuning (Zhang et al., 2023; Li et al., 2023b; Yang et al., 2024b) and external tools (Lu et al., 2023a; Li et al., 2023a) with the development of LLMs. However, these approaches require converting tables into text formats, limiting their applications.

Recently, MLLMs have emerged as a promising direction for table understanding. TableGPT2 (Su et al., 2024a) features a novel table encoder to handle table cell-level information. Pix2Struct (Lee et al., 2023) introduces a unified image-to-text model pretrained on web page screenshots with HTML supervision. PixT3 (Alonso et al., 2024) takes table-to-text tasks as table visual recognition tasks and generates texts. Table-LLaVA (Zheng et al., 2024) introduces a novel multimodal table understanding approach that directly processes table images. However, these approaches do not focus on datasets requiring sophisticated numerical reasoning capabilities.

2.2 Table-based Reasoning and Datasets

Table-based reasoning requires reasoning over both free-form natural language queries and structured tables. Early works either rely on executable languages (*e.g.*, SQL) (Yin et al., 2016; Yu et al., 2018) to capture logical structure in statements. TAPAS (Herzig et al., 2020), and DATER (Ye et al., 2023) encode sentence-table pairs and transform table-based reasoning into question-answering or inference tasks. Existing datasets primarily focus on specific domains like Wikipedia and finance. HybridQA (Chen et al., 2020b) derived from Wikipedia emphasises span lookup, while TAT-QA (Zhu et al., 2021), FinQA (Chen et al., 2021), and DocMath-Eval (Zhao et al., 2024) address numerical reasoning in the financial domain.

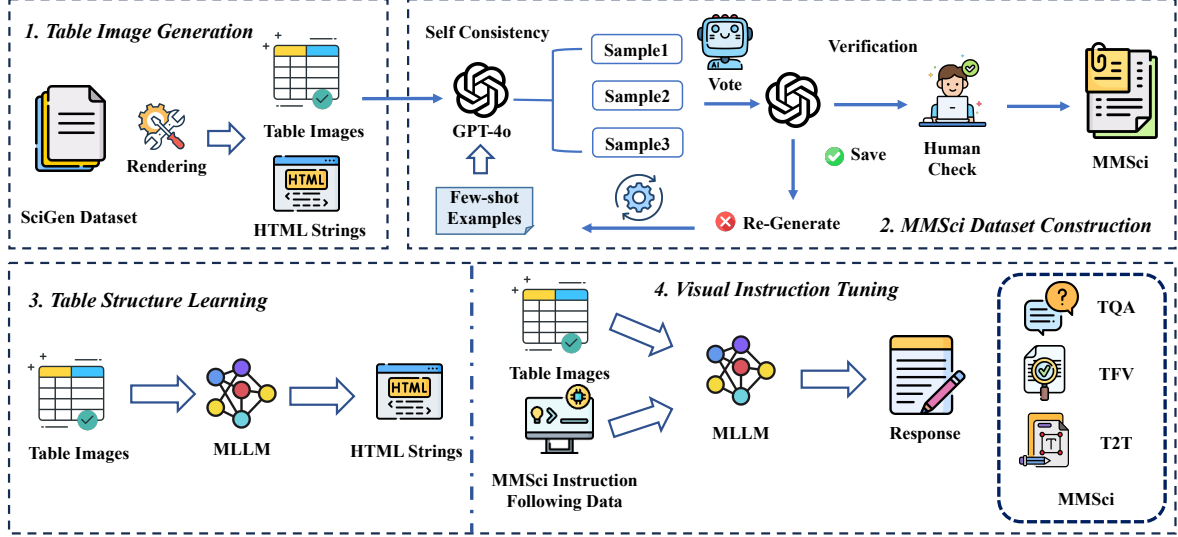


Figure 1: Overview of the proposed framework, which consists of four key stages: (1) Table Image Generation; (2) MMSci Dataset Construction; (3) Table Structure Learning; and (4) Visual Instruction Tuning.

SciGen (Moosavi et al., 2021) introduces a scientific table-to-text generation dataset that requires arithmetic reasoning, but focuses mainly on generation rather than comprehensive reasoning evaluation. However, these datasets have relatively limited reasoning types, significantly differing from real-world scientific table understanding that require numerical computation reasoning. To address this gap, we propose MMSci dataset that combines multiple reasoning types to enhance model performance on complex scientific table understanding tasks.

3 Construction of MMSci Dataset

As shown in Figure 1, the MMSci dataset construction consists of three key components: (1) Data Collection, (2) MMSci-Pre Dataset Construction, and (3) MMSci-Ins and MMSci-Eval Dataset Creation with Numerical Reasoning Augmentation. These components directly correspond to the dataset construction stages in the upper part of our framework.

3.1 Data Collection

To construct MMSci dataset, we focus on scientific tables containing significant numerical values and complex reasoning requirements. We collect raw tabular data from the SciGen dataset (Moosavi et al., 2021), which provides pairs of scientific tables and their corresponding descriptions across computer science research domains. These descriptions naturally require numerical reasoning operations (including addition, subtraction, max/min,

comparison, and division) over table values, making them ideal for our purpose. We transform the original textual tables into high-quality HTML format and then render them into table images while preserving their structural integrity. This process ensures the visual representation maintains the complex layouts and relationships present in the original scientific tables. From this process, we collect 52K image-to-HTML pairs based on tables from the training set and development set of the SciGen dataset.

3.2 MMSci-Pre Dataset Construction

Existing table-based MLLMs (Lee et al., 2023; Alonso et al., 2024; Zheng et al., 2024) demonstrate that generating textual table representations from table images is crucial for aligning visual structure with textual content. Based on our collected image-HTML pairs, we create 52K instruction-following image-to-HTML samples via the `Imgkit`¹ python package. Each sample consists of a table image paired with its corresponding HTML representation. The resulting dataset, MMSci-Pre, contains 52K samples specifically designed for table structure learning.

3.3 MMSci-Ins and MMSci-Eval Dataset Creation

For our instruction tuning and evaluation datasets, we select 12,000 tables from the training set and 1,038 from the testing set of SciGen dataset to

¹<https://pypi.org/project/imgkit/>

create MMSci-Ins and MMSci-Eval, respectively. These datasets focus on complex numerical reasoning tasks requiring multi-step operations including addition, subtraction, comparison, max/min identification, and other mathematical operations. For each table, we employ GPT-4o (OpenAI, 2024) to generate task-specific content across three table-based tasks. For Table Question Answering (TQA), we generate questions paired with step-by-step reasoning processes and final answers. For Table Fact Verification (TFV), we create claims with supporting reasoning steps and verification results (supported, refuted, or not enough information). For Table-to-Text Generation (T2T), we augment existing table-to-text pairs with detailed reasoning steps. To ensure quality, we implement a rigorous verification process. First, we apply self-consistency Chain-of-Thought (CoT) reasoning (Wang et al., 2023) with multiple reasoning paths and voting. Second, we use GPT-4o to validate consistency between reasoning steps and final outputs. Third, we conduct human verification on 40% of generated samples. Finally, we regenerate any identified incorrect samples to maintain dataset quality. This process results in MMSci-Ins with 12K high-quality instruction-tuning samples and MMSci-Eval with 3,114 testing examples. Both datasets maintain a balanced distribution across the three tasks, with each table paired with one sample per task type. Each sample includes detailed step-by-step reasoning processes that enable models to learn both final outputs and the logical progression needed to arrive at conclusions. Detailed dataset quality statistics are provided in Appendix A.2.

4 Experiments

4.1 Model Training

To demonstrate the effectiveness of MMSci dataset, we train two series of MLLM following the architecture of Qwen2-VL-7B-Instruct (Wang et al., 2024) and LLaVA-NeXT-7B (Li et al., 2024).

Model Architectures. Both models follow a three-component design: **Qwen2-VL-7B-Instruct** consists of a Vision Transformer (ViT) (Dosovitskiy, 2020) as the vision tower, a MLP as the vision-language connector, and a Qwen2-7B-Instruct (Yang et al., 2024a) as the language model. **LLaVA-NeXT-7B** uses a pre-trained CLIP model (Radford et al., 2021) as the visual encoder, a MLP connector, and a Vicuna-7B model (Chiang et al., 2023) as the backbone. In both architectures,

the vision encoder processes images into visual features, which are projected into the LLM’s word embedding space via the MLP connector.

We divide the training into two stages:

Table Structure Learning. We use both MMSci-Pre and MMTAB-Pre (Zheng et al., 2024) corpus (202K table image-to-HTML pairs in total) to align visual features with textual representations in different experimental settings as shown in Table 1. Models learn to generate HTML table representations, developing table structure perception capabilities. For LLaVA-NeXT-7B, only the MLP connector parameters are updated during this stage.

Visual Instruction Tuning. We use 12K instruction-following samples from MMSci-Ins to fine-tune the MLLMs while keeping visual encoders frozen. Only the MLP projection layer and LLM weights are updated, focusing on developing instruction-following numerical reasoning capabilities across TQA, TFV, and T2T tasks.

Notably, both models support dynamic input resolutions, addressing a key limitation of existing table MLLMs (Lee et al., 2023; Alonso et al., 2024; Zheng et al., 2024) that require fixed-size input image resolutions (e.g., 336×336). Qwen2-VL achieves this through 2D-RoPE (Su et al., 2024b) to capture two-dimensional positional information of images while LLaVA-NeXT employs a simpler approach of splitting images into grids and encoding them independently. While both Qwen2-VL and LLaVA-NeXT support dynamic input resolutions, our analysis reveals significant performance differences when applied to scientific tables. This suggests that mere technical capability for dynamic resolution is insufficient without proper cross-modal alignment.

4.2 Experimental Settings

Baselines. We select several state-of-the-art MLLMs as our baselines, including GPT-4V (OpenAI, 2023), InternVL-2-76B (Chen et al., 2024), LLaVA-NeXT series (72B/34B/13B/7B) (Li et al., 2024), Qwen-2-VL-Instruct series (72B/7B) (Wang et al., 2024), Table-LLaVA series (13B/7B) (Zheng et al., 2024), Pixtral-12B (Agrawal et al., 2024), Llama-3.2-11B-Vision-Ins. (Meta, 2024), MiniCPM-V-2.6-8B (Yao et al., 2024), and InternVL-2-8B (Chen et al., 2024).

Datasets and Metrics. The held-in evaluation sets in Table 1 include TQA, TFV and T2T tasks of MMSci-Eval. The held-out evaluation sets in Table 2 are from MMTAB-Eval benchmark (Zheng

Models	MMSci-Eval			Held-out	
	TQA Acc.	TFV Acc.	T2T BLEU	TABMWP Acc.	TAT-QA Acc.
Baseline					
GPT-4V (OpenAI, 2023)	53.13	78.01	4.80	60.00	32.50
InternVL-2-76B (Chen et al., 2024)	40.31	62.46	1.79	46.28	6.73
LLaVA-NeXT-72B (Li et al., 2024)	11.75	49.28	1.79	10.69	3.29
Qwen-2-VL-72B-Ins. (Wang et al., 2024)	39.11	64.06	2.83	41.42	17.65
LLaVA-NeXT-34B (Li et al., 2024)	9.73	42.19	2.33	6.96	1.29
LLaVA-NeXT-13B (Li et al., 2024)	2.31	1.83	1.79	1.67	0.43
Table-LLaVA-13B (Zheng et al., 2024)	8.57	51.15	0.03	<u>59.77</u>	15.67
Pixtral-12B (Agrawal et al., 2024)	0.96	5.49	4.12	4.64	7.46
Llama-3.2-11B-Vision-Ins. (Meta, 2024)	1.15	5.85	3.04	7.39	0.37
LLaVA-NeXT-7B (Li et al., 2024)	0.19	0.86	2.99	1.73	0.72
Qwen-2-VL-7B-Ins. (Wang et al., 2024)	25.62	52.79	3.04	34.43	16.19
InternVL-2-8B (Chen et al., 2024)	25.72	44.99	2.64	18.42	7.12
MiniCPM-V-2.6-8B (Yao et al., 2024)	26.58	33.23	0.07	24.30	11.94
Table-LLaVA-7B (Zheng et al., 2024)	7.99	39.30	0.03	57.78	12.82
Ours (LLaVA-NeXT-7B)					
MMSci-Pre (52k) + MMSci-Ins	17.72	57.12	2.93	49.47	10.46
MMTab-Pre (150k) + MMSci-Ins	15.79	56.16	2.88	47.55	8.03
MM-Pre (202k) + MMSci-Ins	23.02	58.57	2.36	49.72	12.27
w/o MM-Pre (202k)	15.22	51.73	2.86	46.24	7.63
Ours (Qwen2-VL-7B-Ins.)					
MMSci-Pre (52k) + MMSci-Ins	41.13	72.92	3.24	49.50	19.68
MMTab-Pre (150k) + MMSci-Ins	40.75	72.73	3.16	49.08	19.30
MM-Pre (202k) + MMSci-Ins	<u>42.10</u>	<u>73.98</u>	<u>3.29</u>	49.96	<u>20.85</u>
w/o MM-Pre (202k)	41.71	70.90	3.29	48.02	20.07

Table 1: Performance comparison on MMSci-Eval and held-out tabular numerical reasoning datasets. MM-Pre (202k) indicates the combination of MMTab-Pre (150k) and MMSci-Pre (52k). w/o MM-Pre represents only training with MMSci-Ins dataset. Best results are in **bold**, second best are underlined.

et al., 2024). TQA contains TABMWP (Lu et al., 2023b), WTQ (Pasupat and Liang, 2015), HiTab (Cheng et al., 2022), TAT-QA (Zhu et al., 2021), and FeTaQA (Nan et al., 2022), where TABMWP and TAT-QA specifically focus on tabular numerical reasoning. TFV contains TabFact (Chen et al., 2020a) and InfoTabs (Gupta et al., 2020), while Table-to-Text (T2T) generation uses HiTab_T2T (Cheng et al., 2022), Rotowire (Wiseman et al., 2017), and WikiBIO (Lebrete et al., 2016). While these datasets contain tables from Wikipedia, financial reports, and government documents, our MMSci-Eval datasets primarily feature scientific tables with numerical values from research papers. We use accuracy and BLEU (Papineni et al., 2002) for TQA, TFV, and T2T benchmarks.

5 Results and Analysis

5.1 Performance on Numerical Reasoning Datasets

The experimental results demonstrate the effectiveness of our proposed approach across various multimodal table understanding tasks. As shown in Table 1, we compare our method with state-of-the-art baselines on both MMSci benchmarks (TQA, TFV, T2T) and held-out tabular numerical reasoning datasets (TABMWP, TAT-QA). Among the baseline models, GPT-4V (OpenAI,

2023) achieves superior performance across all tasks, establishing strong benchmarks with 53.13% accuracy on TQA, 78.01% on TFV, and notably strong generalisation ability on held-out numerical reasoning datasets. Large-scale open-sourced models like InternVL-2-76B (Chen et al., 2024) and Qwen-2-VL-72B (Wang et al., 2024) also demonstrate competitive performance but show relatively weaker generalisation to held-out numerical reasoning datasets.

As for our approaches, with LLaVA-NeXT-7B as the foundation model, we observe that training with MMSci-Pre (52k) dataset demonstrates higher performance (17.72% on TQA, 57.12% on TFV) compared to training with MMTab-Pre (150k) dataset (15.79% on TQA, 56.16% on TFV). The combination of both table structure learning dataset (MM-Pre 202k) further improves performance to 23.02% on TQA and 58.57% on TFV. Notably, our approach shows strong generalisation ability on held-out datasets, achieving 49.72% on TABMWP with the experiment setting of MM-Pre (202k) + MMSci-Ins.

With Qwen2-VL-7B-Instruct as the foundation model, we observe significantly stronger performance across all settings. Training with MMSci-Pre (52k) + MMSci-Ins achieves comparable or better performance (41.13% on TQA, 72.92% on TFV) compared to training with MMTab-Pre (150k) + MMSci-Ins (40.75% on TQA, 72.73%

Method	TQA						TFV			T2T			
	TABMWP	WTQ	HiTab	TAT-QA	FeTaQA	Avg. TQA	TabFact	InfoTabs	Avg. TFV	HiTab_T2T	Rotowire	WikiBIO	Avg. T2T
	Acc.	Acc.	Acc.	Acc.	BLEU	Acc.	Acc.	Acc.	Acc.	BLEU	BLEU	BLEU	BLEU
Baseline													
GPT-4V (OpenAI, 2023)	60.50	48.00	27.50	32.50	11.04	35.91	45.50	<u>65.60</u>	55.55	2.98	4.23	1.94	3.05
Qwen2-VL-7B-Ins. (Wang et al., 2024)	34.44	12.55	3.36	16.19	11.75	15.66	20.28	34.19	27.23	1.90	2.30	2.94	2.38
LLaVA-NeXT-7B (Li et al., 2024)	1.73	0.00	0.00	0.00	1.17	0.58	1.24	1.78	1.51	0.45	1.04	0.67	0.72
Table-LLaVA-7B (Zheng et al., 2024)	57.78	18.43	10.09	12.82	<u>25.60</u>	24.94	<u>59.85</u>	65.26	<u>62.56</u>	<u>9.74</u>	10.46	9.68	9.96
Table-LLaVA-13B (Zheng et al., 2024)	<u>59.77</u>	<u>20.41</u>	<u>10.85</u>	15.67	28.03	<u>26.95</u>	65.00	66.91	65.96	10.40	<u>8.83</u>	<u>9.67</u>	<u>9.63</u>
Ours (LLaVA-NeXT-7B)													
MMSci-Pre (52k) + MMSci-Ins	8.76	3.22	0.63	0.39	5.99	3.80	35.78	25.37	30.57	1.57	1.10	1.78	1.48
MMTab-Pre (150k) + MMSci-Ins	9.00	2.62	0.63	0.26	7.23	3.95	36.22	26.91	31.56	1.64	0.84	1.57	1.35
MM-Pre (202k) + MMSci-Ins	10.66	4.83	0.82	0.65	9.39	5.27	39.63	27.63	33.63	1.13	0.83	1.90	1.29
w/o MM-Pre (202k)	9.69	2.74	0.19	0.39	6.84	3.97	31.72	23.80	27.76	1.69	0.79	1.53	1.34
Ours (Qwen2-VL-7B-Ins.)													
MMSci-Pre (52k) + MMSci-Ins	49.51	18.74	4.95	19.69	12.89	21.15	37.93	45.33	41.63	0.75	2.81	2.69	2.08
MMTab-Pre (150k) + MMSci-Ins	49.09	18.95	4.63	19.30	9.77	20.35	40.00	46.56	43.28	0.91	1.26	2.89	1.69
MM-Pre (202k) + MMSci-Ins	46.97	19.73	4.38	<u>20.85</u>	12.34	20.85	39.99	45.96	42.97	0.96	1.32	2.60	1.63
w/o MM-Pre (202k)	48.02	18.67	5.33	20.08	12.58	20.94	33.53	44.93	39.23	0.71	2.76	2.70	2.06

Table 2: Performance comparison on MMTAB held-out datasets. Best results are in bold, second best are underlined.

Models	MMSci-Eval			Held-out	
	TQA	TFV	T2T	TABMWP	TAT-QA
Ours (LLaVA-NeXT-7B)					
MMSci-Pre (52k) + MMSci-Ins	17.72	57.12	2.93	49.47	10.46
w/o Reasoning	10.75	42.73	2.16	42.50	7.68
MMTab-Pre (150k) + MMSci-Ins	15.79	56.16	2.88	43.55	8.03
w/o Reasoning	9.58	50.31	1.93	42.50	7.42
MM-Pre (202k) + MMSci-Ins	23.02	58.57	2.36	49.72	12.27
w/o Reasoning	12.73	45.21	2.16	46.50	19.68
w/o MM-Pre (202k)	15.22	51.73	2.86	46.24	7.63
w/o Reasoning	9.43	42.31	2.36	45.50	8.39
Ours (Qwen2-VL-7B-Ins.)					
MMSci-Pre (52k) + MMSci-Ins	41.13	72.92	3.24	49.50	19.68
w/o Reasoning	35.06	66.47	3.14	44.08	16.72
MMTab-Pre (150k) + MMSci-Ins	40.75	72.73	3.16	49.08	19.30
w/o Reasoning	34.48	66.28	2.27	43.97	16.07
MM-Pre (202k) + MMSci-Ins	42.10	73.98	3.29	49.96	20.85
w/o Reasoning	35.45	67.43	1.97	46.34	17.68
w/o MM-Pre (202k)	41.71	70.90	3.29	48.02	20.07
w/o Reasoning	34.44	62.90	3.18	44.60	14.68

Table 3: Ablation study results for reasoning steps on MMSci-Eval and held-out datasets.

on TFV), despite using only one-third of the table structure learning data. The experiment setting of training with MM-Pre (202k) + MMSci-Ins achieves the best performance with 42.10% accuracy on TQA and 73.98% on TFV, while also demonstrating strong generalisation ability on held-out numerical reasoning datasets (49.96% on TABMWP and 20.85% on TAT-QA).

These results demonstrate that our proposed MMSci-Pre dataset with 52K scientific domain-specific data is more effective than MMTAB-Pre with 150K general-domain data, highlighting the importance of data quality over quantity. Furthermore, Qwen2-VL-7B-Instruct consistently outperforms LLaVA-NeXT-7B across all experimental settings, suggesting its stronger capability in table understanding and numerical reasoning tasks. Besides, our approach shows strong generalisation to held-out tabular numerical reasoning datasets, demonstrating enhanced general ability in multi-modal table understanding and reasoning.

While the absolute performance (42.10% on TQA) may appear modest, it represents substantial improvement for the challenging task of scientific table reasoning. Similar to early work in other complex domains, these results establish important baselines that future research can build upon.

5.2 Performance on Held-out MMTAB Benchmarks

The experimental results in Table 2 also demonstrate the effectiveness and generalisation ability of our proposed approach across various held-out MMTAB benchmark. GPT-4V (OpenAI, 2023) show strong performance across all tasks, achieving 35.91% average accuracy on TQA, 55.55% on TFV, and 3.05 BLEU on T2T. Table-LLaVA models, which are specifically trained on MMTAB-Ins dataset, demonstrate competitive performance. Table-LLaVA-13B achieves strong results on TFV (65.96% average accuracy) and T2T (9.63 BLEU) while Table-LLaVA-7B shows robust performance on TABMWP (57.78%).

As for our approaches, with LLaVA-NeXT-7B as the foundation model, we observe that training with MMSci-Pre (52k) and MMSci-Ins, despite not being trained on MMTAB-Ins dataset (Zheng et al., 2024), demonstrates promising generalisation ability. The MMSci-Pre (52k) + MMSci-Ins combination achieves 3.80% average accuracy on TQA and 30.57% on TFV with only scientific domain data. The combination of both table structure learning datasets (MM-Pre 202k) further improves performance across all metrics, reaching 5.27% on TQA and 33.63% on TFV. As for Qwen2-VL-7B-Instruct as the foundation model, we observe significantly stronger generalisation capa-

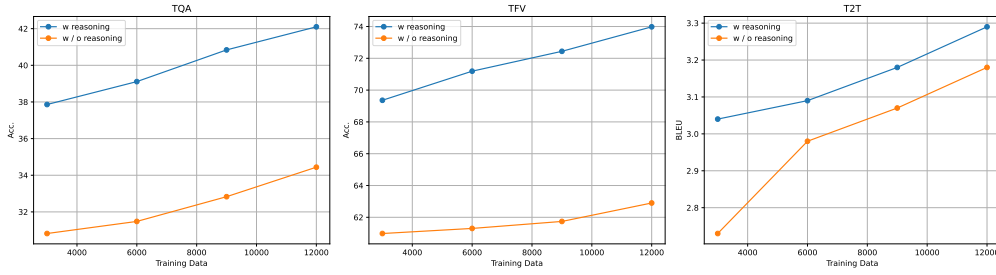


Figure 2: Performance scaling with increasing instruction tuning data size on three MMSci tasks.

bility. MMSci-Pre (52k) + MMSci-Ins combination achieves 21.15% average accuracy on TQA and 41.63% on TFV, demonstrating strong zero-shot transfer to MMTab benchmark despite using only scientific domain data (MMSci dataset). This performance is particularly impressive when compared to MMTab-Pre (150k) + MMSci-Ins combination, which uses three times more image-to-HTML data. Even without any table structure learning (w/o MM-Pre), our approach achieves competitive results, highlighting the effectiveness of our MMSci-Ins instruction tuning dataset.

These results empirically demonstrate our MMSci-Pre dataset with 52K scientific domain-specific data achieves comparable or better performance than MMTab-Pre with 150K general-domain data in MMTab held-out benchmark, highlighting the importance of scientific domain-specific tables. Even without MMTab table structure learning data, our approach demonstrates strong generalisation ability, particularly evident in the performance of MMSci-Pre (52k) + MMSci-Ins and w/o MM-Pre experiment settings.

5.3 Ablation Study on Reasoning Steps

We evaluate the effectiveness of reasoning steps across different experiment configurations. As shown in Table 3, Qwen2-VL-7B-Instruct demonstrates superior performance across all configurations. Without reasoning steps, the model training with MMSci-Pre (52k) + MMSci-Ins achieves better results than that with MMTab-Pre (150k) + MMSci-Ins, highlighting the importance of domain-specific table structure learning over data quantity. Adding reasoning steps consistently improves performance across all metrics, with the model reaching its peak performance under the MM-Pre (202k) + MMSci-Ins experiment configuration. Similar trends are observed in LLaVA-NeXT-7B, though with lower overall performance. These patterns extend to held-out tabular numer-

ical reasoning datasets, where both models show strong generalisation capabilities with reasoning steps, especially on numerical reasoning tasks like TABMWP and TAT-QA. The results demonstrate that a smaller amount of scientific domain-specific table structure learning data, combined with explicit reasoning steps, can be more effective than larger-scale general domain table structure learning.

5.4 Impact of Training Data Size

As shown in Figure 2, we compare performance of MLLMs training with MM-Pre (202k) + MMSci-Ins across three MMSci tasks (TQA, TFV, T2T) with instruction tuning data (MMSci-Ins dataset) size increasing from 3K to 12K samples. The findings demonstrate consistent advantages of incorporating reasoning steps across all data scales. Models trained with reasoning steps maintain substantial performance advantages across all tasks (7-8% for TQA, 8-10% for TFV, 0.3-0.4 BLEU for T2T). While both variants benefit from increased training data, models with reasoning steps show stronger scaling behavior, particularly in TQA and TFV tasks. The persistent performance gap across all data sizes suggests that reasoning steps provide fundamental improvements in model learning that cannot be simply achieved through increased training data alone.

5.5 Representational Alignment Analysis

In this section, we conduct an in-depth analysis to assess the language-vision alignment from the perspective of the representation space. This analysis aims to provide further insights into the observed variations in model performance, particularly in the context of scientific multimodal table understanding and reasoning tasks.

Preliminaries. We formalise MLLMs within the framework of an *unembedding-embedding* architecture. In this framework, the unembedding stage

Models	Cycle KNN	Mutual KNN	Les KNN	CKA	CKNNA	SVCCA	Edit KNN
<i>Unembedding stage: ImageNet(Concepts)</i>							
Random	0.02761	0.01257	0.52355	0.08614	0.00714	0.12425	0.00019
Qwen2-VL-7B-Ins.	0.68110	<u>0.03486</u>	<u>1.28153</u>	<u>0.08856</u>	0.03067	0.14318	0.00112
Llama3.2-11B-Vision-Ins.	<u>0.08608</u>	0.04205	1.52788	0.06079	0.01403	0.11651	0.00061
LLaVA-NeXT-7B	0.57173	0.02077	0.81645	0.08024	0.01577	0.13240	0.00037
Phi3.5-Vision-Ins.	0.02761	0.01257	0.52355	0.08614	0.00714	0.12118	0.00019
InternVL2-8B	0.08175	0.01637	0.72495	0.09185	0.00062	0.12148	0.00044
<i>Unembedding stage: Wikipedia Caption (short descriptive sentences)</i>							
Qwen2-VL-7B-Ins.	<u>0.49414</u>	0.06855	2.05078	0.08876	0.04093	0.20229	0.00175
Llama3.2-11B-Vision-Ins.	0.31347	0.03623	1.29980	0.00968	0.00779	<u>0.22120</u>	0.00050
LLaVA-NeXT-7B	0.57813	0.03935	1.36523	<u>0.07933</u>	<u>0.03998</u>	0.23114	0.00082
Phi3.5-Vision-Ins.	0.04980	0.03027	1.14843	0.01669	0.03890	0.18183	0.00066
InternVL2-8B	0.36914	0.04132	<u>1.55761</u>	0.04732	0.01658	0.21739	0.00093
<i>Embedding stage: MMSci T2T tasks (table to text description).</i>							
Qwen2-VL-7B-Ins.	0.38631	0.06726	<u>2.03660</u>	0.19318	<u>0.05514</u>	0.38461	0.00183
Llama3.2-11B-Vision-Ins.	0.31310	0.02200	0.84007	1.73979e-8	0.03208	0.08180	0.00026
LLaVA-NeXT-7B	0.38246	0.04514	1.49325	0.15203	0.06673	<u>0.28857</u>	0.00109
Phi3.5-Vision-Ins.	<u>0.38053</u>	<u>0.06712</u>	2.12909	<u>0.16121</u>	0.03688	0.26982	<u>0.00127</u>
InternVL2-8B	0.36512	0.04651	1.56647	0.04230	0.02675	0.11876	0.00096

Table 4: Kernel alignment analysis. The representation for each sample is the averaged token embeddings. The best two values are shown in **bold** and underlined.

is responsible for learning transformations between observations (e.g., text, vision) and latent spaces through encoders, while the embedding stage captures the complex interactions among latent variables within the latent space of LLMs’ hidden layers. Each stage serves distinct functions and yields representations with different properties (Park et al., 2024). Consequently, by focusing on each stage independently, we can have a systematical evaluation of model behaviours in representation spaces. To assess the representational alignment between vision-language modalities at each stage, we next measure the geometrical similarity between them via the *kernel*.

Kernels, characterising the distance metrics between points in a representation space, are commonly used to assess vector space (Huh et al., 2024). Typically, the more similarity between two kernels derived from different spaces (text or vision) indicates a higher degree of alignment between those modality spaces. This similarity can be quantified via *kernel-alignment metrics*, such as Centered Kernel Distance (CKA) (Kornblith et al., 2019). For more information about kernel-alignment metrics used in the experiment, we refer to Huh et al. (2024) for a deep understanding.

Quantitative evaluation. For the unembedding stage, we specifically choose two language-vision datasets: ImageNet (Deng et al., 2009) and Wikipedia Caption (WIT) (Srinivasan et al., 2021). We randomly select 2048 samples from each dataset. These datasets offer varying levels of fine granularity in language-vision alignment, enabling a comprehensive assessment of representational performance. As illustrated in Table 4, we can ob-

serve that the Qwen2-VL-7B-Instruct can generally outperform other baselines on both datasets, indicating it has better fine-grained alignment between language and vision. In the embedding stage, we evaluate alignment on the MMSci T2T task. Since some models do not support single-modality input, we utilise a reference language model (e.g., openllama- 7B (Geng and Liu, 2023)) as the text encoder and MLLMs as the image encoder with prompt “please describe the table”. Alignment is measured based on the output embedding from the last hidden layer. As shown in Table 4, Qwen2-VL-7B-Instruct outperforms the other models, demonstrating its superior language-vision alignment capability. This segment of the experiment demonstrates that the Qwen2-VL-7B-Instruct model exhibits superior language-vision alignment within the representation space. This finding is consistent with the cross-modal consistency analysis presented in Appendix C.1, where we evaluate different table information modalities as inputs to MLLMs and assess their cross-modal consistency (i.e., the proportion of identical predictions) on TQA and TFV tasks.

6 Conclusion

In this paper, we introduce a comprehensive framework for multimodal scientific table understanding and reasoning with dynamic input image resolutions. Experimental results validate our framework’s effectiveness across different model architectures, showing consistent improvements in both general table understanding and numerical reasoning capabilities, with strong generalisation to held-out datasets.

Limitations

While this work advances scientific multimodal table understanding and reasoning, several limitations remain for future research. First, our framework primarily focuses on scientific tables containing numerical values, while other types of scientific tables (e.g., qualitative comparison tables, methodology tables) are not extensively covered. Second, though our framework demonstrates strong performance on numerical reasoning tasks, the current approach may still struggle with complex statistical analyses and domain-specific mathematical notations that are common in scientific literature. Third, while our models support dynamic input resolutions, processing extremely large tables with dense information remains challenging due to computational constraints and potential information loss during visual encoding.

Ethical Statement

The MMSci datasets are constructed from publicly available scientific papers and their associated tables, primarily sourced from open-access repositories and academic databases with appropriate licenses. All table images are generated through automated scripts from the original scientific papers, maintaining their integrity while ensuring proper attribution. The instruction tuning samples are created based on the original scientific context, preserving the academic nature of the source material. Our framework is designed to assist in scientific research by improving the accessibility and understanding of tabular data in academic literature. We anticipate that this work will contribute positively to the research community by facilitating more efficient analysis of scientific publications. The code and datasets are made publicly available for research purposes, promoting transparency and reproducibility in the field of multimodal scientific table understanding.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv: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.

Pei Chen, Soumajyoti Sarkar, Leonard Lausen, Balasubramaniam Srinivasan, Sheng Zha, Ruihong Huang, and George Karypis. 2023. Hytrel: Hypergraph-enhanced tabular data representation learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wenhu Chen. 2023. **Large language models are few(1)-shot table reasoners**. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. 2020a. **Tabfact: A large-scale dataset for table-based fact verification**.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Intervl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. **Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality**.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. **Imagenet: A large-scale hierarchical image database**. In *2009 IEEE Conference*

712	on Computer Vision and Pattern Recognition, pages	767
713	248–255.	768
714	Alexey Dosovitskiy. 2020. An image is worth 16x16	769
715	words: Transformers for image recognition at scale.	770
716	<i>arXiv preprint arXiv:2010.11929</i> .	771
717	Xinyang Geng and Hao Liu. 2023. Openllama: An open	772
718	reproduction of llama .	773
719	Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek	774
720	Srikumar. 2020. INFOTABS: Inference on tables	775
721	as semi-structured data . In <i>Proceedings of the 58th</i>	776
722	<i>Annual Meeting of the Association for Computational</i>	777
723	<i>Linguistics</i> , pages 2309–2324, Online. Association	
724	for Computational Linguistics.	
725	Jonathan Herzig, Pawel Krzysztof Nowak, Thomas	
726	Müller, Francesco Piccinno, and Julian Martin Eisen-	
727	schlos. 2020. Tapas: Weakly supervised table parsing	
728	via pre-training . In <i>Proceedings of the 58th Annual</i>	
729	<i>Meeting of the Association for Computational Lin-</i>	
730	<i>guistics (ACL)</i> , pages 4320–4333.	
731	Minyoung Huh, Brian Cheung, Tongzhou Wang, and	
732	Phillip Isola. 2024. Position: The platonic represen-	
733	tation hypothesis . In <i>Proceedings of the 41st Inter-</i>	
734	<i>national Conference on Machine Learning</i> , volume	
735	235 of <i>Proceedings of Machine Learning Research</i> ,	
736	pages 20617–20642. PMLR.	
737	Simon Kornblith, Mohammad Norouzi, Honglak Lee,	
738	and Geoffrey Hinton. 2019. Similarity of neural	
739	network representations revisited. In <i>International</i>	
740	<i>conference on machine learning</i> , pages 3519–3529.	
741	PMLR.	
742	Rémi Lebrete, David Grangier, and Michael Auli. 2016.	
743	Neural text generation from structured data with ap-	
744	plication to the biography domain . In <i>Proceedings of</i>	
745	<i>the 2016 Conference on Empirical Methods in Natu-</i>	
746	<i>ral Language Processing</i> , pages 1203–1213, Austin,	
747	Texas. Association for Computational Linguistics.	
748	Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu,	
749	Fangyu Liu, Julian Eisenschlos, Urvashi Khandel-	
750	wal, Peter Shaw, Ming-Wei Chang, and Kristina	
751	Toutanova. 2023. Pix2Struct: Screenshot Parsing	
752	as Pretraining for Visual Language Understanding .	
753	ArXiv:2210.03347 [cs].	
754	Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Ren-	
755	rui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and	
756	Chunyan Li. 2024. Llava-next: Stronger llms super-	
757	charge multimodal capabilities in the wild .	
758	Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and	
759	Zhaoxiang Zhang. 2023a. Sheetcopilot: Bringing	
760	software productivity to the next level through large	
761	language models . In <i>Thirty-seventh Conference on</i>	
762	<i>Neural Information Processing Systems</i> .	
763	Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge,	
764	Haidong Zhang, Danielle Rifinski Fainman, Dong-	
765	mei Zhang, and Surajit Chaudhuri. 2023b. Table-gpt:	
766	Table-tuned gpt for diverse table tasks .	
	Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi	767
	Lin, Weizhu Chen, and Jian-Guang Lou. 2022.	768
	TAPEX: Table pre-training via learning a neural SQL	769
	executor . In <i>International Conference on Learning</i>	770
	<i>Representations</i> .	771
	Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-	772
	Wei Chang, Ying Nian Wu, Song-Chun Zhu, and	773
	Jianfeng Gao. 2023a. Chameleon: Plug-and-play	774
	compositional reasoning with large language mod-	775
	els. In <i>The 37th Conference on Neural Information</i>	776
	<i>Processing Systems (NeurIPS)</i> .	777
	Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu,	778
	Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark,	779
	and Ashwin Kalyan. 2023b. Dynamic prompt learn-	780
	ing via policy gradient for semi-structured mathe-	781
	matical reasoning. In <i>International Conference on</i>	782
	<i>Learning Representations (ICLR)</i> .	783
	Meta. 2024. Llama 3.2: Pushing the boundaries of	784
	vision and language for edge and mobile devices .	785
	Accessed: 2024-12-06.	786
	Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth,	787
	and Iryna Gurevych. 2021. Scigen: a dataset for	788
	reasoning-aware text generation from scientific ta-	789
	bles. In <i>Thirty-fifth Conference on Neural Informa-</i>	790
	<i>tion Processing Systems Datasets and Benchmarks</i>	791
	<i>Track (Round 2)</i> .	792
	Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria	793
	Lin, Neha Verma, Rui Zhang, Wojciech Kryściński,	794
	Hailey Schoelkopf, Riley Kong, Xiangru Tang,	795
	Mutethia Mutuma, Ben Rosand, Isabel Trindade,	796
	Renusree Bandaru, Jacob Cunningham, Caiming	797
	Xiong, and Dragomir Radev. 2022. Fetaqa: Free-	798
	form table question answering. <i>Transactions of the</i>	799
	<i>Association for Computational Linguistics</i> , 10:35–49.	800
	OpenAI. 2023. Gpt-4v. https://openai.com/index/	801
	gpt-4v-system-card/ . Accessed: 2023-02-09,	802
	2023-02-11, 2023-02-12.	803
	OpenAI. 2024. Hello gpt-4o. https://openai.com/	804
	index/hello-gpt-4o/ . Accessed: 2024-02-09,	805
	2024-02-11, 2024-02-12.	806
	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	807
	Jing Zhu. 2002. Bleu: a method for automatic eval-	808
	uation of machine translation. In <i>Annual Meeting of</i>	809
	<i>the Association for Computational Linguistics</i> .	810
	Kiho Park, Yo Joong Choe, and Victor Veitch. 2024.	811
	The linear representation hypothesis and the geom-	812
	etry of large language models . In <i>Proceedings of</i>	813
	<i>the 41st International Conference on Machine Learn-</i>	814
	<i>ing</i> , volume 235 of <i>Proceedings of Machine Learning</i>	815
	<i>Research</i> , pages 39643–39666. PMLR.	816
	Panupong Pasupat and Percy Liang. 2015. Composi-	817
	tional semantic parsing on semi-structured tables . In	818
	<i>Proceedings of the 53rd Annual Meeting of the As-</i>	819
	<i>sociation for Computational Linguistics and the 7th</i>	820
	<i>International Joint Conference on Natural Language</i>	821

822	<i>Processing (Volume 1: Long Papers)</i> , pages 1470–	Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang,	879
823	1480, Beijing, China. Association for Computational	Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 techni-	880
824	Linguistics.	cal report .	881
825	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya	Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao,	882
826	Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-	and Chenghua Lin. 2024b. Effective distillation of	883
827	try, Amanda Askell, Pamela Mishkin, Jack Clark,	table-based reasoning ability from llms. In <i>Pro-</i>	884
828	Gretchen Krueger, and Ilya Sutskever. 2021. Learn-	<i>ceedings of the 2024 Joint International Conference</i>	885
829	ing transferable visual models from natural language	<i>on Computational Linguistics, Language Resources</i>	886
830	supervision .	<i>and Evaluation (LREC-COLING 2024)</i> , pages 5538–	887
831	Krishna Srinivasan, Karthik Raman, Jiecao Chen,	5550.	888
832	Michael Bendersky, and Marc Najork. 2021. Wit:	Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang,	889
833	Wikipedia-based image text dataset for multimodal	Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li,	890
834	multilingual machine learning . <i>ArXiv preprint,</i>	Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v:	891
835	<i>abs/2103.01913</i> .	A gpt-4v level mllm on your phone . <i>ArXiv preprint,</i>	892
836	Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou,	<i>abs/2408.01800</i> .	893
837	Ga Zhang, Guangcheng Zhu, Haobo Wang, Haokai	Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei	894
838	Xu, Hao Chen, Haoze Li, et al. 2024a. Tablegpt2: A	Huang, and Yongbin Li. 2023. Large language mod-	895
839	large multimodal model with tabular data integration.	els are versatile decomposers: Decomposing evi-	896
840	<i>arXiv preprint arXiv:2411.02059</i> .	dence and questions for table-based reasoning . In	897
841	Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan,	<i>Proceedings of the 46th International ACM Confer-</i>	898
842	Wen Bo, and Yunfeng Liu. 2024b. Roformer: En-	<i>ence on Research and Development in Information</i>	899
843	hanced transformer with rotary position embedding.	<i>Retrieval (SIGIR)</i> , pages 174–184.	900
844	<i>Neurocomputing</i> , 568:127063.	Pengcheng Yin, Zhengdong Lu, Hang Li, and Ben Kao.	901
845	Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and	2016. Neural enquirer: Learning to query tables in	902
846	Dongmei Zhang. 2023. Gpt4table: Can large lan-	natural language . In <i>Proceedings of the 25th Inter-</i>	903
847	guage models understand structured table data? a	<i>national Joint Conference on Artificial Intelligence</i>	904
848	benchmark and empirical study .	<i>(IJCAI)</i> , pages 2308–2314.	905
849	Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhi-	Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga,	906
850	hao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin	Dongxu Wang, Zifan Li, James Ma, Irene Li,	907
851	Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhanc-	Qingning Yao, Shanelle Roman, Zilin Zhang, and	908
852	ing vision-language model’s perception of the world	Dragomir R. Radev. 2018. Spider: A large-scale	909
853	at any resolution . <i>ArXiv preprint</i> , abs/2409.12191.	human-labeled dataset for complex and cross-domain	910
854	Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc	semantic parsing and text-to-sql task . In <i>Proceedings</i>	911
855	Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery,	<i>of the 2018 Conference on Empirical Methods in</i>	912
856	and Denny Zhou. 2023. Self-Consistency Improves	<i>Natural Language Processing (EMNLP)</i> , pages 3911–	913
857	Chain of Thought Reasoning in Language Models .	3921.	914
858	<i>ArXiv:2203.11171 [cs]</i> .	Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun.	915
859	Sam Wiseman, Stuart Shieber, and Alexander Rush.	2023. Tablellama: Towards open large generalist	916
860	2017. Challenges in data-to-document generation .	models for tables .	917
861	In <i>Proceedings of the 2017 Conference on Empiri-</i>	Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi,	918
862	<i>cal Methods in Natural Language Processing</i> , pages	Linyong Nan, Lyuhao Chen, Yixin Liu, Xian-	919
863	2253–2263, Copenhagen, Denmark. Association for	gru Tang, Rui Zhang, and Arman Cohan. 2024.	920
864	Computational Linguistics.	DocMath-eval: Evaluating math reasoning capabili-	921
865	An Yang, Baosong Yang, Binyuan Hui, Bo Zheng,	ties of LLMs in understanding long and specialized	922
866	Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan	documents . In <i>Proceedings of the 62nd Annual Meet-</i>	923
867	Li, Dayiheng Liu, Fei Huang, Guanting Dong, Hao-	<i>ing of the Association for Computational Linguis-</i>	924
868	ran Wei, Huan Lin, Jialong Tang, Jialin Wang,	<i>tics (Volume 1: Long Papers)</i> , pages 16103–16120,	925
869	Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin	Bangkok, Thailand. Association for Computational	926
870	Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai,	Linguistics.	927
871	Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-	Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She,	928
872	qin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni,	Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024.	929
873	Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize	Multimodal Table Understanding . In <i>Proceedings</i>	930
874	Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan,	<i>of the 62nd Annual Meeting of the Association for</i>	931
875	Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge,	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,	932
876	Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren,	pages 9102–9124, Bangkok, Thailand. Association	933
877	Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing	for Computational Linguistics.	934
878	Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan,		

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. [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)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Details about MMSci

A.1 Datasets Statistics

Table 5 presents the distribution of reasoning types in our MMSci-Eval dataset. The most common type is addition (21.1%), followed by subtraction (15.3%) and max/min operations (15.7%). Division and comparison operations also appear frequently (14.2% and 13.7% respectively). More complex operations like ranking (9.6%) and look-up (8.9%) occur less frequently, while domain knowledge calculations are rare (1.5%).

The average number of reasoning steps varies significantly across types, with subtraction requiring the most steps (4.1) and look-up operations requiring the fewest (1.5). This variation reflects the inherent complexity of different mathematical operations and their application to tabular data. Notably, even seemingly simple operations like addition require multiple steps (2.8) on average, indicating the complexity of reasoning with tabular scientific data.

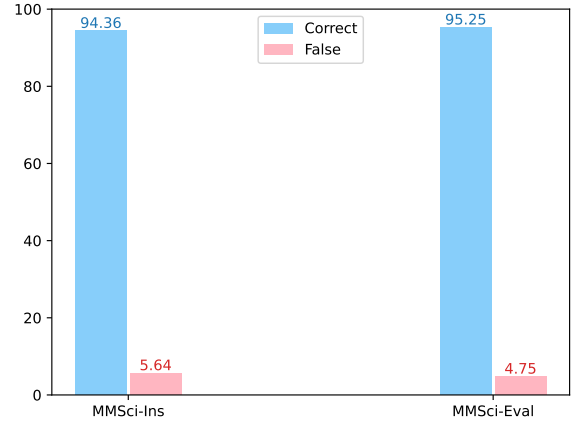


Figure 3: Evaluation of generated data of MMSci-Ins and MMSci-Eval dataset. Correct refers to the data verified correctly by human annotators.

A.2 Dataset Quality Control

To ensure data quality, we conduct a rigorous human verification process for both MMSci-Ins and MMSci-Eval datasets. For MMSci-Ins, we manually verify 40% of the generated samples, achieving a high accuracy rate of 94.36%. For MMSci-Eval, given its critical role as a benchmark, we carefully examine all 3,114 generated samples and achieve an accuracy of 95.25%. For any identified incorrect samples, we employ GPT-4o to regenerate them following the same self-consistency voting mechanism, followed by another round of both automatic

Reasoning Type	Description	Avg. Reasoning Step	Prop.%
Add	Calculate the sum between numbers	2.8	21.1
Comparison	Comparison of values	2.1	13.7
Domain Knowledge Calculation	Calculations need domain knowledge	2.2	1.5
Divide	Perform division between numbers	3.4	14.2
Look Up	Search for cells in tables	1.5	8.9
Max/Min	Retrieve the maximum or minimum number	3.2	15.7
Ranking	Arranges items in a specific order	2.4	9.6
Subtract	Perform subtraction between numbers	4.1	15.3

Table 5: The reasoning types, descriptions, average reasoning step, and proportion in our dataset.

and manual verification to ensure quality. This iterative process ensures the reliability and correctness of our datasets for both training and evaluation purposes.

A.3 Prompt for Generating Data

The prompt for MMSci-Ins and MMSci-Pre data generation is shown in Table 7.

B Experimental Settings

Implementation Details. Both models follow a three-component design. Qwen2-VL-7B-Instruct consists of a Vision Transformer (ViT) (Dosovitskiy, 2020) as the vision tower, a MLP as the vision-language connector, and Qwen2 (Wang et al., 2024) as the language model. LLaVA-NeXT-7B uses a pre-trained CLIP model (Radford et al., 2021) as the visual encoder, a MLP connector, and Vicuna-7B (Chiang et al., 2023) as the backbone. In both architectures, the vision encoder processes images into visual features, which are projected into the LLM’s word embedding space via the MLP connector.

Training Details. All experiments are conducted on 4×A100 80GB GPUs using LoRA with rank 64 and sequence length 4096. For table structure learning, LLaVA-NeXT-7B requires 15 hours for MMTAB-Pre (150k), 3 hours for MMSci-Pre (52k), and 20 hours for combined training (one epoch). Qwen2-VL-7B takes 15 hours, 8 hours, and 19 hours respectively. The instruction tuning stage requires approximately 1 hour for 4 epochs with 12k samples for both models.

C More Experimental Results and Analysis

C.1 Vision-Language Consistency Analysis

We evaluate the cross-modal consistency of different MLLMs by comparing their performance

when processing table information through different modalities. For each model, we test with both table images (image modality) and their textual representations (text modality), measuring both task performance (Acc.) and cross-modal consistency (Cons.).

Qwen2-VL-7B-Instruct demonstrates superior cross-modal alignment, achieving the highest consistency scores on both TQA (60.40%) and TFV (72.48%) tasks. Notably, it maintains strong performance across both modalities, with image-based accuracy (TQA: 39.11%, TFV: 52.79%) consistently outperforming text-based results (TQA: 21.65%, TFV: 50.10%). This suggests robust integration of visual and textual understanding capabilities.

Other models show varying degrees of modality gap. MiniCPM-V-2.6-8B and InternVL-2-8B achieve moderate consistency (48.78% and 50.89% on TQA), while models like LLaVA-NeXT-7B and Pixtral-12B show significant disparities between modalities, resulting in lower consistency scores. These results highlight the challenge of maintaining consistent reasoning capabilities across different input modalities in table understanding tasks.

D Dataset Examples and Case Study

D.1 Dataset Examples

The training examples of MMSci-Pre Dataset are shown in Figure 4. The examples of MMSci-Eval are shown Figure 6.

D.2 Case Study

The case study of TQA and TFV tasks are shown in Figure 6 and Figure 7, respectively

Model	Modal	TQA		TFV	
		Acc.	Consis.	Acc.	Consis.
Qwen2-VL-7B-Ins. (Wang et al., 2024)	Text Image	21.65 39.11	60.40	50.10 52.79	72.48
LLaVA-NeXT-7B (Li et al., 2024)	Text Image	3.17 0.19	14.81	2.03 0.86	23.65
MiniCPM-V-2.6-8B (Yao et al., 2024)	Text Image	21.11 26.58	48.78	30.82 33.23	38.53
InternVL-2-8B (Chen et al., 2024)	Text Image	19.84 25.72	50.89	42.87 44.99	36.42
Pixtral-12B (Agrawal et al., 2024)	Text Image	1.44 0.96	16.52	4.43 5.49	29.88
Llama-3.2-11B-Vision-Ins. (Meta, 2024)	Text Image	3.24 1.15	15.71	6.96 5.85	20.40

Table 6: Vision-language consistency evaluation across different MLLMs. Consistency scores measure the percentage of identical responses between modalities, indicating the model’s cross-modal alignment.

Task	Input: Please review this table image and return a text representation of the table in HTML format.						Output:																																																																																								
Table Image-to-HTML	Table Image <table> <tr> <th>Model</th><th>Target Corpus</th><th>Transfer BERT</th><th>Transfer Filler</th><th>Transfer Role</th><th>Baseline Acc. (%)</th><th>Fine-tuned Acc. (%)</th><th>Gain (%)</th></tr> <tr> <td>BERT</td><td>QNLI</td><td>True</td><td>—</td><td>—</td><td>91.60</td><td>91.27</td><td>−0.33</td></tr> <tr> <td>BERT</td><td>QQP</td><td>True</td><td>—</td><td>—</td><td>91.45</td><td>91.12</td><td>−0.33</td></tr> <tr> <td>BERT</td><td>RTE</td><td>True</td><td>—</td><td>—</td><td>71.12</td><td>73.65</td><td>+2.53</td></tr> <tr> <td>BERT</td><td>SNLI</td><td>True</td><td>—</td><td>—</td><td>90.45</td><td>90.69</td><td>+0.24</td></tr> <tr> <td>BERT</td><td>SST</td><td>True</td><td>—</td><td>—</td><td>93.23</td><td>92.78</td><td>−0.45</td></tr> <tr> <td>HUBERT (Transformer)</td><td>QNLI</td><td>True</td><td>True</td><td>False</td><td>90.56</td><td>91.16</td><td>+0.60</td></tr> <tr> <td>HUBERT (Transformer)</td><td>QQP</td><td>False</td><td>False</td><td>True</td><td>90.81</td><td>91.42</td><td>+0.61</td></tr> <tr> <td>HUBERT (Transformer)</td><td>RTE</td><td>True</td><td>True</td><td>True</td><td>61.73</td><td>74.01</td><td>+12.28</td></tr> <tr> <td>HUBERT (Transformer)</td><td>SNLI</td><td>True</td><td>False</td><td>True</td><td>90.66</td><td>91.36</td><td>+0.70</td></tr> <tr> <td>HUBERT (Transformer)</td><td>SST</td><td>True</td><td>False</td><td>True</td><td>91.28</td><td>92.43</td><td>+1.15</td></tr> </table> <p>Table 3: Transfer learning results for GLUE tasks. The source corpus is MNLI. Baseline accuracy is when Transfer BERT, Filler, and Role are all False, equivalent to no transfer. Fine-tuned accuracy is the best accuracy among all possible transfer options.</p>						Model	Target Corpus	Transfer BERT	Transfer Filler	Transfer Role	Baseline Acc. (%)	Fine-tuned Acc. (%)	Gain (%)	BERT	QNLI	True	—	—	91.60	91.27	−0.33	BERT	QQP	True	—	—	91.45	91.12	−0.33	BERT	RTE	True	—	—	71.12	73.65	+2.53	BERT	SNLI	True	—	—	90.45	90.69	+0.24	BERT	SST	True	—	—	93.23	92.78	−0.45	HUBERT (Transformer)	QNLI	True	True	False	90.56	91.16	+0.60	HUBERT (Transformer)	QQP	False	False	True	90.81	91.42	+0.61	HUBERT (Transformer)	RTE	True	True	True	61.73	74.01	+12.28	HUBERT (Transformer)	SNLI	True	False	True	90.66	91.36	+0.70	HUBERT (Transformer)	SST	True	False	True	91.28	92.43	+1.15	<pre> <table border="1" cellspacing="0"> <tr> <th>Model</th> <th>Target Corpus</th> <th>Transfer BERT</th> <th>Transfer Filler</th> <th>Transfer Role</th> <th>Baseline Acc. (%)</th> <th>Fine-tuned Acc. (%)</th> <th>Gain (%)</th> </tr> <tr> <td>BERT</td> <td>QNLI</td> <td>True</td> <td>—</td> <td>—</td> <td>91.60</td> <td>91.27</td> <td>−0.33</td> </tr> <tr> <td>BERT</td> <td>QQP</td> <td>True</td> <td>—</td> <td>—</td> <td>91.45</td> <td>91.12</td> <td>−0.33</td> </tr> <tr> <td>BERT</td> <td>RTE</td> <td>True</td> <td>—</td> <td>—</td> <td>71.12</td> <td>73.65</td> <td>+2.53</td> </tr> <tr> <td>BERT</td> <td>SNLI</td> <td>True</td> <td>—</td> <td>—</td> <td>90.45</td> <td>90.69</td> <td>+0.24</td> </tr> <tr> <td>BERT</td> <td>SST</td> <td>True</td> <td>—</td> <td>—</td> <td>93.23</td> <td>92.78</td> <td>−0.45</td> </tr> <tr> <td>HUBERT (Transformer)</td> <td>QNLI</td> <td>True</td> <td>True</td> <td>False</td> <td>90.56</td> <td>91.16</td> <td>+0.60</td> </tr> <tr> <td>HUBERT (Transformer)</td> <td>QQP</td> <td>False</td> <td>False</td> <td>True</td> <td>90.81</td> <td>91.42</td> <td>+0.61</td> </tr> <tr> <td>HUBERT (Transformer)</td> <td>RTE</td> <td>True</td> <td>True</td> <td>True</td> <td>61.73</td> <td>74.01</td> <td>+12.28</td> </tr> <tr> <td>HUBERT (Transformer)</td> <td>SNLI</td> <td>True</td> <td>False</td> <td>True</td> <td>90.66</td> <td>91.36</td> <td>+0.70</td> </tr> <tr> <td>HUBERT (Transformer)</td> <td>SST</td> <td>True</td> <td>False</td> <td>True</td> <td>91.28</td> <td>92.43</td> <td>+1.15</td> </tr> </table> </pre>
Model	Target Corpus	Transfer BERT	Transfer Filler	Transfer Role	Baseline Acc. (%)	Fine-tuned Acc. (%)	Gain (%)																																																																																								
BERT	QNLI	True	—	—	91.60	91.27	−0.33																																																																																								
BERT	QQP	True	—	—	91.45	91.12	−0.33																																																																																								
BERT	RTE	True	—	—	71.12	73.65	+2.53																																																																																								
BERT	SNLI	True	—	—	90.45	90.69	+0.24																																																																																								
BERT	SST	True	—	—	93.23	92.78	−0.45																																																																																								
HUBERT (Transformer)	QNLI	True	True	False	90.56	91.16	+0.60																																																																																								
HUBERT (Transformer)	QQP	False	False	True	90.81	91.42	+0.61																																																																																								
HUBERT (Transformer)	RTE	True	True	True	61.73	74.01	+12.28																																																																																								
HUBERT (Transformer)	SNLI	True	False	True	90.66	91.36	+0.70																																																																																								
HUBERT (Transformer)	SST	True	False	True	91.28	92.43	+1.15																																																																																								

Figure 4: MMSci-Pre Dataset example

The prompt for Generating data

You are given a table image and a description: {description}.

1.For the Table to Text (T2T) task, come up with a one to two sentence length succinct multi-hop reasoning step of the description.

Write your results as 'T2T Reasoning:' and then the succinct reasoning step.

2.For the Table Question Answering (TQA) task, come up with a question and answer with multi-hop reasoning step.

The question and answer must be based on the table image and description.

Write your results as 'TQA Question:' and then the question and 'TQA Reasoning:' and then the reasoning step and 'TQA Answer:' and then the answer.

When generating 'TQA Question:', make sure it is a single question that requires reasoning based on the table.

When generating 'TQA Answer:', provide the final answer in the JSON structure, using the format "answer": "<YOUR ANSWER>"

Make sure the answer only contains one entity, such as 'So, the answer is "answer": "23".'

3.For the Table Fact Checking (TFV) task, come up with a statement and answer with multi-hop reasoning step.

The statement and answer must be based on the table image and description. The table 'supports' or 'refutes' the statement. The statement should be considered 'not enough info' if it may or may not be true.

Write your results as 'TFV Statement:' and then the statement and 'TFV Reasoning:' and then the reasoning step and 'TFV Answer:' and then the answer.

Make sure the answer only contains one entity, such as 'Thus, the answer is "answer": "supports".'

When generating 'TFV Answer:', provide the final answer in the JSON structure, using the format "answer": "<YOUR ANSWER>"

Fill the result into JSON format without any other words:

```
"T2T Reasoning": "<YOUR T2T REASONING>",  
"TQA Question": "<YOUR TQA QUESTION>",  
"TQA Reasoning": "<YOUR TQA REASONING>",  
"TQA Answer": "<YOUR TQA ANSWER>",  
"TFV Statement": "<YOUR TFV STATEMENT>",  
"TFV Reasoning": "<YOUR TFV REASONING>",  
"TFV Answer": "<YOUR TFV ANSWER>"
```

Examples:

```
{TQA Examples}  
{TFV Examples}  
{T2T Examples}
```

Table 7: The prompts for generating the questions, reasoning steps, and answers or claims of MMSci-Ins and MMSci-Eval datasets.

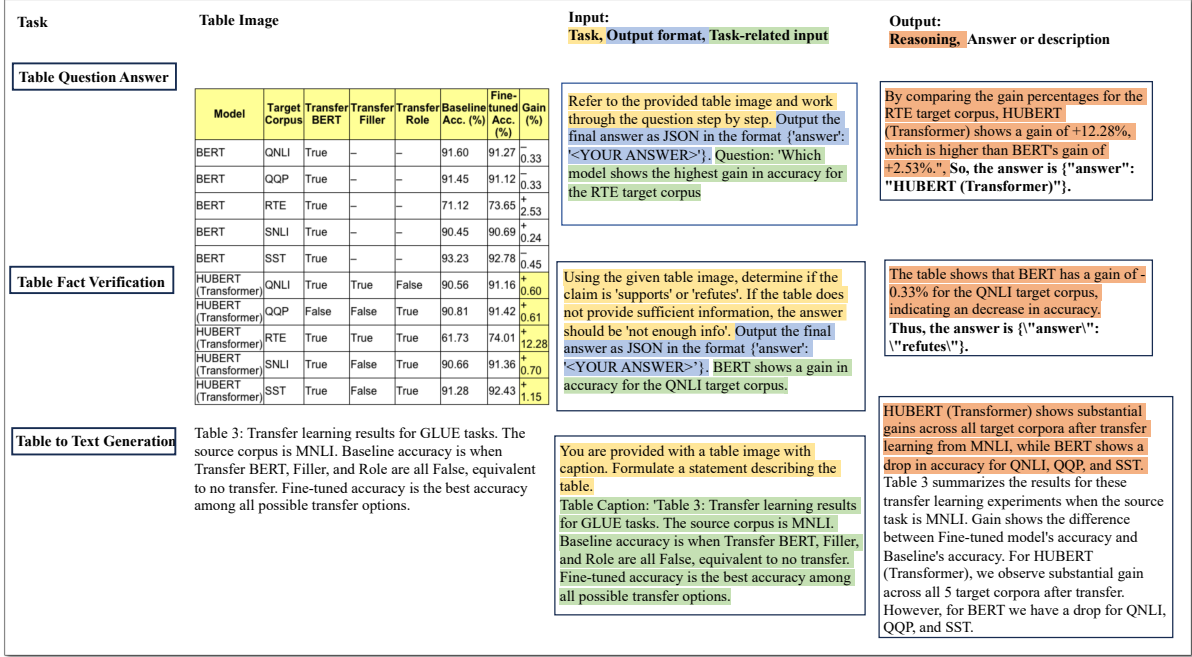


Figure 5: MMSci-Ins and MMSci-Eval Dataset example

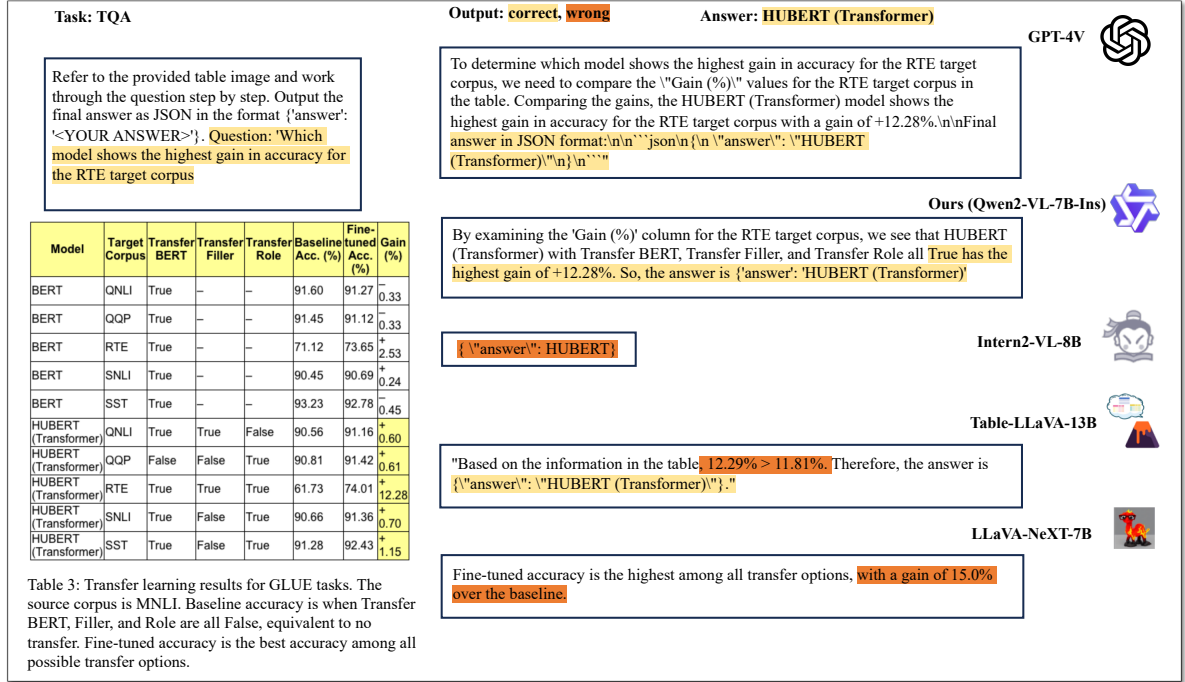


Figure 6: A case example of TQA task

