Modality-Swap Distillation: Rendering Textual Reasoning into Visual Supervision

Boammani Aser Lompo *
École de Technologie Supérieure
Montreal, Canada
boammani.lompo.1@ens.etsmtl.ca

Marc Haraoui *
mharaoui.pro@gmail.com

Abstract

Visual reasoning over structured data such as tables is a critical capability for modern vision-language models (VLMs), yet current benchmarks remain limited in scale, diversity, or reasoning depth, especially when it comes to rendered table images. Addressing this gap, we introduce **Visual-TableQA**, a large-scale, open-domain multimodal dataset specifically designed to evaluate and enhance visual reasoning over complex tabular data. Our generation pipeline is **modular, scalable, and fully autonomous**, involving multiple reasoning LLMs collaborating across distinct roles: generation, validation, and inspiration. **Visual-TableQA** comprises 2.5k richly structured LaTeX-rendered tables and 9k reasoning-intensive QA pairs, all produced at a cost of under $100. To promote diversity and creativity, our pipeline performs **multi-model collaborative data generation** via **cross-model prompting ('inspiration')** and LLM-jury filtering. Stronger models seed layouts and topics that weaker models elaborate, collectively distilling diverse reasoning patterns and visual structures into the dataset. Empirical results show that models fine-tuned on **Visual-TableQA** generalize robustly to external benchmarks, outperforming several proprietary models despite the dataset's synthetic nature. The full pipeline and resources are publicly available in our GitHub repository.

1 Introduction

Vision-language models (VLMs) have significantly advanced in recent years, achieving remarkable performance in various tasks involving visual and textual inputs. Despite these advancements, complex reasoning tasks, especially those requiring deep comprehension of tabular data structures, continue to pose significant challenges. Table complexity can manifest in various ways, including structural layout, information density, and the diversity of visual components such as the integration of diagrams. The more complex a table is, the more it lends itself to challenging reasoning tasks, requiring advanced cognitive abilities to extract relevant information and perform multi-step logical analysis. For example, the table in Figure 1 exemplifies this complexity through its use of multirow cells, integrated diagrams, and color encoding. Answering the question requires the VLM to interpret information across all cells and perform a sequence of reasoning steps.

Existing table-based QA datasets predominantly fall into two categories: *(i)* those represented purely in textual format—such as WikiTableQuestions Pasupat & Liang (2015), HybridQA Chen et al. (2020b), and AIT-QA Katsis et al. (2022)—which bypass the challenges of visual layout interpretation; and *(ii)* those that lack diversity in visual layouts, visual complexity, and reasoning depth due to being domain-specific (e.g., TAT-DQA Zhu et al. (2022)), or having standardized queries (e.g., TableVQA-Bench Kim et al. (2024)), or highly technical in nature (e.g., Table-VQA Tom Agonnoude (2024)).

*Equal contribution, random ordering.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundations of Reasoning in Language Models.

Comparison of Lighting Setups in Photography by Climate Zone

Climate Zone	Natural Light Quality	Recommended Setup	Key Modifier	Color Temperature (K)	Common Challenges	Sample Subjects
Tropical	Intense, direct	Diffuser + Reflector	White umbrella	5600	Harsh shadows	Portraits, Flora
Arid	Bright, high contrast	Scrim + Fill Flash	Silver reflector	6000	Glare, overexposure	Wildlife, Landscape
Temperate	Variable, soft	Softbox + Bounce	Softbox	5500	Changing weather	Fashion, Product
Continental	Clear, crisp	Strobe + Grid	Grid spot	5700	Color cast	Architecture, Street
Polar	Low, blue-tinted	LED Panel + Gel	Blue gel	6500	Low light, cold	Documentary, Ice forms

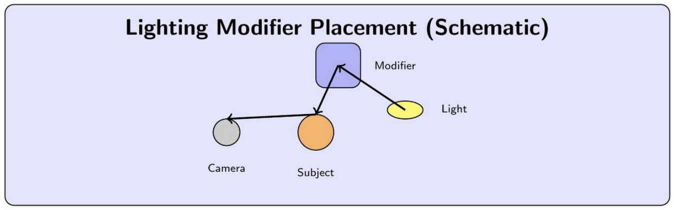


Figure 1: Sample question in our benchmark

This second datasets category typically rely on a limited set of layout templates and involve relatively simple visual tasks or basic QA scenarios, falling short of the complexity required for thorough evaluation and advancement of reasoning capabilities. More recent efforts—such as ChartQA Masry et al., ReachQA He et al., and MATH-Vision Wang et al. (2024b)—have aimed to address the need for open-domain coverage, incorporating more diverse visual features, varied question types, and deeper reasoning challenges. However, these datasets primarily focus on charts and function plots, overlooking tables—and with them, an entire dimension of informational structure and layout diversity. An extensive comparison of diverse chart and table datasets is provided in Table 1.

Inspired by ReachQA's Code-as-Intermediary Translation (CIT)—a technique that translates chart images into textual representations while faithfully preserving visual features—we introduce **Visual-TableQA**, a novel synthetic, multimodal, and open-domain dataset tailored to enhance reasoning capabilities through complex table-based question-answering tasks. Visual-TableQA capitalizes on the ability of reasoning-oriented LLMs to generate intricate LaTeX tables, thus significantly reducing costs and eliminating the need for extensive manual annotations. This modality-swap makes it possible for LLMs to invest their textual reasoning ability into visual image in order to improve visual understanding and reasoning. Visual-TableQA emphasizes structural reasoning over domain knowledge. Each entry couples a rendered table image with a complex, visually grounded reasoning task. Tasks require interpreting visual layout cues such as cell alignment, hierarchical headers, merged cells, or embedded symbolic content—emulating real-world documents where visual context is essential for correct interpretation. The dataset contains 2.5k reasoning-intensive tables and 9k QA pairs crafted to assess both information extraction and multi-step reasoning capabilities, all generated at a cost of under \$100. The entire dataset has been validated using a committee of high-performing reasoning LLMs, the ROSCOE step by step reasoning score Golovneva et al., and a sample of 800 QA pairs has undergone manual verification by human annotators. In contrast to previous synthetic datasets, Visual-TableQA is less guided in its generation process, allowing for more diversity and creativity in both table complexity (e.g., structural layout, information density, visual component variety) and the design of QA pairs explicitly crafted to challenge visual reasoning skills. We evaluated a broad range of VLMs, from lightweight models to state-of-the-art architectures, and benchmarked their performance against existing datasets. The results show that most VLMs continue to struggle with table understanding.

In sum, our main contributions are:*(i)* a high-quality, visually diverse, and open-domain dataset for table-based reasoning; *(ii)* an LLM-driven, low-cost generation pipeline using cross-model inspiration; *(iii)* an empirical analysis comparing Visual-TableQA to existing table and chart datasets; *(iv)* an extensive evaluation of open and proprietary VLMs, showing performance gains after fine-tuning. Our dataset and code are publicly available at `https://github.com/AI-4-Everyone/Visual-TableQA`.

Table 1: Comparison of existing chart and table datasets across data, Q&A, and dataset properties. Abbreviations: Repr=Representation, Vis= Visual, Comp= Complexity, Temp = Template, Refer = Reference, Rat = Rational, Synth= Synthetic, Scal = Scalable. Cells marked with ▲indicate mixed attributes (e.g., partially template-based; scalable Q&A but non-scalable chart data)

	Data Properties				Q&A Properties			Dataset Properties		
Datasets	# Layouts/ # Topics	Type	Data Repr.	Vis. Comp.	Temp. Free	Vis. Re-fer.	Rat. An-not.	Synth.	#Samples / #QA	Scal.
WikiTableQuestions (Pasupat & Liang, 2015)	–	Table	Text	✗	✗	✗	✗	✗	2.1k/22k	✗
HybridQA (Chen et al., 2020b)	–	Table	Text	✗	✓	✗	✗	✗	13k/70k	✗
AIT-QA (Katsis et al., 2022)	–/1	Table	Text	✗	✓	✗	✗	✗	116/515	✗
TAT-DQA (Zhu et al., 2022)	–/1	▲	Image	✗	✓	✓	✓	✗	2.5k/16.5k	✗
Table-VQA (Tom Agonnoude, 2024)	–/–	Table	Image	✗	✓	✓	✓	✓	16.4k/82.3k	–
TableVQA-Bench (Kim et al., 2024)	11/4	Table	Image	✗	✓	✓	✗	▲	894/1.5k	▲
ChartQA (Masry et al.)	3/15	Chart	Image	✗	✓	✓	✗	✗	21.9k/32.7k	✗
DocVQA (Mathew et al., 2020)	20/5	▲	Image	✗	✓	✓	✗	✗	12.7k/50k	✗
MultiModalQA (Talmor et al., 2021)	16/∞	▲	Image	✗	✗	✓	✗	▲	29,918	✗
MATH-Vision (Wang et al., 2024b)	–/16	▲	Image	✓	✓	✓	✗	✗	3k/3k	✗
REACHQA (He et al.)	32/∞	Chart	Image	✓	✓	✓	✓	✓	3.7k/22k	✓
Visual-TableQA (ours)	/∞	Table	Image	✓	✓	✓	✓	✓	2.5k/ 9k	✓

2 Visual-TableQA Dataset

Unlike previous datasets that rely heavily on textual input or handcrafted annotations, Visual-TableQA leverages a scalable generation pipeline rooted in LaTeX-rendered table images, automated reasoning task creation, and LLM-based evaluation. This strategy enables high diversity and reasoning depth while keeping annotation costs minimal, totaling under $100 using a combination of open-access APIs and limited usage tiers. In this section, we describe our LaTeX-based table encoding 2.1, the data generation pipeline 2.2, and the quality assurance process 2.3.

2.1 Modality-Swap: Table Representation in LaTeX

Our approach is inspired by He et al., which demonstrated that state-of-the-art VLMs can reason about visual content even in the absence of explicit visual input. Building on this insight, and leveraging the strong coding capabilities of reasoning-oriented language models across multiple programming languages, we chose to use an intermediate representation of tables in LaTeX rather than directly generating rendered table images. This strategy enables the generation of complex visual tables as compact LaTeX code—typically around 100 lines per table—drastically reducing the cost of generation by minimizing the number of output tokens required in API calls. We refer to this as modality-swap: LLMs invest their textual reasoning ability (programming and analysing programs) into visual images in order to improve visual understanding and reasoning. An ablation study of the modality-swap assumption is reported in Table 2. We observe that most state-of-the-art models perform better when the same query is provided in textual form rather than as a visual input. This suggests both room for improvement in visual reasoning and the potential for transferring textual reasoning capabilities to the visual modality.

Our observations (Table 3) align with those of Kale & Nadadur (2025), who reported that LLMs struggle with LaTeX generation—particularly as task complexity increases, leading to a notable drop in accuracy. Table 3 presents the performance of various models in generating LaTeX tables that compile without errors.

2.2 Data Generation Pipeline

This section provides a detailed description of the generation pipeline. Figure 2 gives an overview of the whole process.

Table 2: Model performance on the test sets of four benchmarks: ChartQA, ReachQA, MATH-Vision, Visual-TableQA, and Visual-TableQA-CIT. Visual-TableQA-CIT is the variant of our dataset where tables are represented in LaTeX code form rather than as rendered images. The ReachQA score is reported as the average across its two evaluation splits: *Reasoning* and *Recognition*. The values in blue are from our own evaluation using the LLM jury, while the remaining values are taken from model authors or official leaderboards/model cards. When a fine-tuned model achieves better performance, the result is annotated with ↑; if the performance worsens, it is marked with ↓. The best performance for each model variants and task is in **bold**.

Models	**ChartQA**	**ReachQA**	**MATH-Vision$_{FULL}$**	**Visual-TableQA**	**Visual-TableQA-CIT**
			Baseline		
Human	–	74.85	68.82	–	–
			Proprietary VLMs		
GPT-4o	85.7	53.25	30.39	78.5	91.0
GPT-4o mini	77.52	40.35	28.85	67.0	80.27
Gemini 2.5 Flash	84.64	56.97	41.3	85.72	92.3
Gemini 2.5 Pro	85.73	61.87	**73.3**	78.62(86.63)	86.27
Claude 3.5 Sonnet	**90.8***	**63**	32.76	**82.46**	**88.5**
			Open-Source VLMs		
Llama 4 Maverick 17B-128E Instruct	85.3*	47.98 (49.75)	**45.89**	80.75	87.0
Mistral Small 3.1 24B Instruct	86.24*	42.45	32.45	73.2	80.25
Qwen2.5-VL-32B-Instruct	**79.75**	**49.5**	38.1	**80.45**	**81.93**
Qwen2.5-VL-7B-Instruct	87.3*	49.23	25.1	71.35	–
			Finetuned VLMs		
Qwen2.5-VL-7B-Instruct + Visual-TableQA	**84.52**↑	**60.95** ↑	**49.77** ↑	**82.98** ↑	N/A
Qwen2.5-VL-7B-Instruct + ReachQA	77.59 ↓	55.75 ↑	48.57 ↑	60.68 (56.13) ↓	N/A

* Performance metrics are measured using *Relaxed Accuracy*, which allows for small numerical deviations in the predicted answers. We assume that this accuracy inflates the actual accuracy by at least 5%. This margin is subtracted when selecting the best-performing results, which are shown in **bold**.

Table 3: Percentage of successful LaTeX compilations for various models. Each accuracy is computed from at least 500 generated samples. The Adjust column indicates the level of manual correction needed to make the table look good: Low means minimal or no adjustments, Medium corresponds to 3–5 required fixes, and High indicates more than 5 adjustments were necessary. The tables generated by DeepSeek-R1-Distill-Qwen-32B never compiled.

Model	**Acc. (%)**	**Adjust**	**Model**	**Acc. (%)**	**Adjust**
Llama 4 Maverick 17B-128E Instruct Meta AI (2025)	69	High	DeepSeek-R1-Distill-Qwen-32B DeepSeek-AI (2025)	0.0	–
Gemini 2.0 Flash Google (2025a)	65.7	Low	DeepSeek-R1T-Chimera TNG Technology Consulting GmbH (2025)	43.4	Medium
Gemini 2.5 Flash Google (2025b)	43	Medium	Claude Sonnet 4 Anthropic (2025)	56	Low
Gemini 2.5 Pro Google (2025c)	19.6	Low	Claude 3.5 Haiku Anthropic (2024)	64.4	Low
GPT-4.1 OpenAI (2025a)	41.5	Low	Grok 3 Beta xAI (2025)	47.3	Low
Qwen3-30B-A3B Qwen Team (2025a)	69.4	Low	Reka Flash 3 Reka AI (2025)	19.3	Medium
Qwen-QwQ-32B Qwen Team (2025b)	38.2	Low			

Seed Tables and Topics Collection: The first step involves collecting a diverse set of table layouts to serve as inspiration for LLMs during the generation process. We explored various sources, including scientific journals, financial report databases, online newspapers, and table design galleries. Our search included both table and diagram images to introduce greater visual and structural complexity into the dataset. We selected 20 representative images (Figure 6a) and passed them to a visual language model, VLM-0 (`GPT-o3` OpenAI (2025)), to generate accurate LaTeX representations. In parallel, we used LLM-0 (`GPT-4o`) OpenAI (2024) to generate a list of 5,000 distinct topic prompts. These initial table samples and topics serve as the first layer of inspiration for subsequent LLM generations—though the pool of inspirations expands automatically, as detailed in Section 2.2. For reproducibility, all resources are publicly available in our GitHub repository.

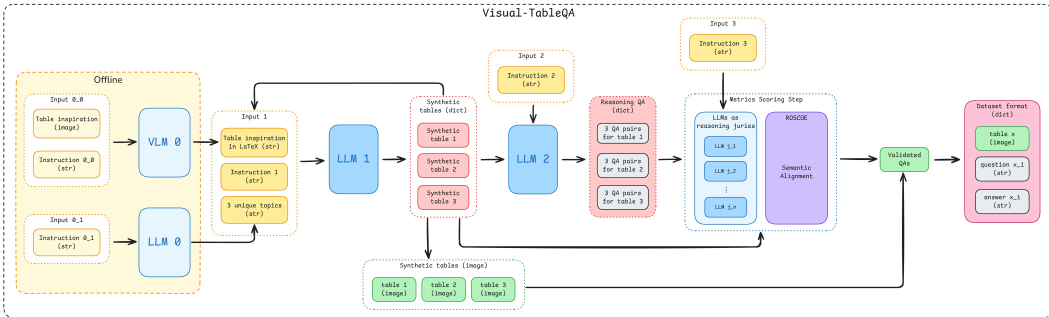


Figure 2: Overview of the full pipeline architecture of Visual-TableQA. A subset of initial table images is first converted to LaTeX using a visual language model (VLM-0). The resulting LaTeX code, along with topic prompts, is then passed to a language model (LLM-1) to generate new, diverse tables. These newly generated tables are used as inputs for further iterations of table generation. All generated tables are then submitted to a second language model (LLM-2), which produces corresponding question-answer pairs. Finally, the QA pairs are evaluated by a jury of high-performing LLMs, and their quality is assessed using the ROSCOE score.

Table Generation: For each iteration, we randomly select an LLM-1 from the models short-list presented in Table 3. The model receives one table sample from our pool and three topics randomly selected from the topic list, all delivered through a single instruction prompt. The output from LLM-1 is returned as a JSON file containing three newly generated LaTeX-formatted tables in plain text, each corresponding to one of the provided topics. We require that the generated tables be inspired by the input table but include substantial layout variations and, when appropriate, additional data to enhance complexity. The resulting LaTeX code is then compiled using standard LaTeX compilation stack (pdflatex + pdf2image), and cropped to produce high-resolution table images. A human reviewer then inspects the table and makes adjustments to the LaTeX code if necessary. The prompt used for generation are provided in Figure 7.

Evolving Layouts through Iterations: A subset of the generated tables is manually selected to enrich the pool of table inspirations. This feedback loop encourages the emergence of increasingly complex and diverse layouts by amplifying visual variations and enabling cross-model inspiration across different LLM-1s over successive iterations. This process is facilitated by the fact that LLMs differ in architecture and tend to focus on distinct structural and stylistic aspects of tables. As a result, combining inspirations across models leads to highly diversified and creative layout types. We refer to this phenomenon as **cross-model prompting ('inspiration')**.

QA Generation: Next, for each generated table, we randomly select a model, denoted LLM-2, from the same list of models in Table 3 to generate three QA pairs. The model receives the table in LaTeX format and is instructed to produce questions that require multi-step reasoning, pattern recognition, and symbolic interpretation. For instance, the sample in Figure 1 illustrates how the questions extend beyond basic information extraction, requiring interpretative reasoning to identify patterns within the presented data. We do not fact-check the generated tables; as a result, some table content may be non-factual. While this is important to consider when using the dataset for training, it can be beneficial, as it encourages models to rely on reasoning rather than prior knowledge.

2.3 Quality Assurance

To ensure the validity of the tables and QA pairs, a panel of independent LLMs—serving as a reasoning jury—evaluates each table and its associated QA pairs by providing binary correctness judgments. The evaluation is based on four criteria: *(i)* the generated document is a valid table and is relevant to the given topic; *(ii)* the table and any associated figures are coherent and meaningful; *(iii)* the question is fully grounded in the table, requiring no external knowledge; and *(iv)* the answer is completely supported by the table content. If any of these four criteria are not met, the corresponding table and its QA pairs are discarded. The LLM jury includes `Mistral-large`, `Deepseek-v3.1`, `Gemini-2.5-pro`, `GPT-4.1`, and `Deepcogito-v2`—models chosen for their strong reasoning abilities. Final acceptance is determined via majority vote across the jury. The prompt used is provided in Figure 9.

The next step involved computing the ROSCOE reasoning scores as introduced in Golovneva et al.. These metrics assess the coherence, logical soundness, and contextual grounding of step-by-step generated rationales. The ROSCOE framework encompasses thirteen evaluation criteria, which we report in Table 7 along with their corresponding values computed over our dataset. The results indicate near-perfect alignment with the expected directionality of each metric, supporting the overall quality of the generated reasoning chains.

Test Set Construction and Human Evaluation: The dataset was divided into three subsets: training, validation, and testing. To prevent data leakage, all entries {table, question, answer} derived from a single table were assigned to the same subset. The testing set was also used for human evaluation. Two human annotators—each holding at least a Master's degree and with prior experience in data annotation—were hired to evaluate the quality of 800 QA pairs. Each QA pair was assessed for validity and rated on a scale from 1 to 5. Overall, 92% of the evaluated QA pairs received a rating of at least 4 stars from both annotators.

3 Experiments

3.1 Benchmark Comparison

Evaluated Benchmarks and Model Selection: We evaluate a range of state-of-the-art reasoning VLMs on Visual-TableQA and compare their performance across three other benchmarks focused on table and chart-based visual question answering: ChartQA Masry et al., ReachQA He et al., and MATH-Vision Wang et al. (2024b). Our model selection includes powerful proprietary models such as `GPT-4o`, `GPT-4o Mini` OpenAI (2025b), `Gemini 2.5 Flash`, `Gemini 2.5 Pro`, and `Claude 3.5 Sonnet`, as well as open-source models like `LLaMA 4 Maverick 17B-128E Instruct`, `Mistral Small 3.1 24B Instruct` Mistral AI (2025), `Qwen2.5-VL-32B-Instruct` Chen et al. (2024a), `Qwen2.5-VL-7B-Instruct` Team (2025), `LLaVA-Next-Llama3-8B` Li et al. (2024), `MiniCPM-V2.5-Llama3` Yao et al. (2024), and `InternVL2-8B` Chen et al. (2024b). Where performance metrics were available, we did not re-evaluate models on these datasets; instead, we report the results published in the original papers, official leaderboards, or model cards. For all other cases, we carefully fine-tuned and evaluated the models following the instructions provided in their respective official GitHub repositories.

Evaluation Protocol: All models are evaluated on the test sets of the four selected datasets. Each model receives image-question pairs, formatted within a unified prompt that includes a system message tailored to elicit the model's reasoning capabilities (Section G.1). For the Visual-TableQA dataset, we additionally construct a variant in which data is provided not as rendered images but in LaTeX code format. This textual-code version is referred to as Visual-TableQA-CIT.

For `LLaVA-Next-Llama3-8B`, `MiniCPM-V2.5-Llama3`, `InternVL2-8B`, and `Qwen2.5-VL-7B-Instruct`, we conducted two supervised fine-tuning (SFT) experiments: *(i)* using the ReachQA training split (denoted as `Model_Name + ReachQA`) and *(ii)* using the Visual-TableQA training split (denoted as `Model_Name + Visual-TableQA`). We applied Low-Rank Adapters (LoRA) Hu et al. to all linear layers, following the SFT setup and hyperparameters described in the He et al. GitHub repository when possible (Section H) in order to make a fair comparison. The fine-tuning phase for all models was limited to one epoch to ensure consistency and reduce overfitting. Exceptionally, we adopted a custom two-phase LoRA fine-tuning strategy for `Qwen2.5-VL-7B-Instruct` (see Section H), as this model was not included in the evaluation of He et al., and to better accommodate the relatively small size of our dataset.

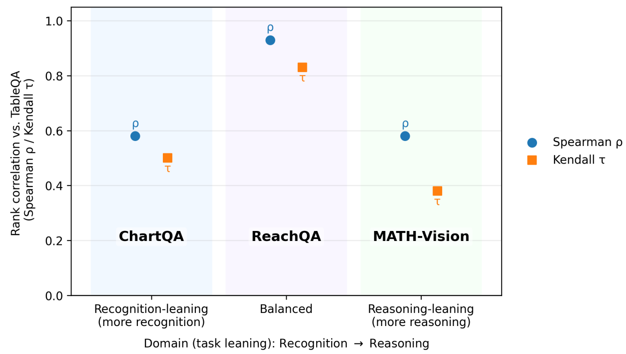
All models are allocated a maximum of 5,000 tokens during inference to accommodate extended chain-of-thought reasoning. Model responses are evaluated using the same jury of high-performing VLMs and majority-vote protocol as described in Section 2.3. The jury confidence score, computed as the ratio of the highest vote count to the total jury size, averages above 0.87 for all models and all datasets. In addition, evaluations are run twice, to ensure reproducibility.

3.2 Experimentation Results

The average models accuracies are displayed in Table 2. These results reveal that:

Visual-TableQA Effectively Evaluates Visual Reasoning Capabilities: Model performances on

Figure 3: Correlation of model rankings on Visual-TableQA with those on three established datasets—ChartQA (recognition-focused), ReachQA (balanced), and MATH-Vision (reasoning-focused)—using Spearman's ρ and Kendall's τ metrics. Higher values indicate stronger alignment in model performance trends. Visual-TableQA shows strong correlation with ReachQA, suggesting it effectively balances both visual recognition and reasoning, while its weaker correlation with ChartQA and MATH-Vision highlights its unique position as a comprehensive visual reasoning benchmark.



Visual-TableQA follow similar trends to those observed on real-world, human-annotated datasets such as ChartQA and MATH-Vision, suggesting that synthetic datasets can effectively evaluate reasoning capabilities. A direct comparison between Visual-TableQA and its textual variant, Visual-TableQA-CIT, shows a notable performance gap: on average, models perform $+6.26\%$ better on Visual-TableQA-CIT. This highlights the added challenge posed by the image-based format in Visual-TableQA, demonstrating its effectiveness at testing visual reasoning over purely textual input.

To further validate Visual-TableQA as a reasoning benchmark, we compared model rankings across datasets. For each dataset, we extracted the models (except the fine-tuned ones) performance rankings and compared them to the rankings on Visual-TableQA using two correlation measures: *(i)* Spearman's ρ Lee Rodgers & Nicewander (1988): Captures monotonic consistency in rankings (regardless of exact scores); *(ii)* Kendall's τ Kendall (1948): Measures the fraction of concordant vs. discordant ranking pairs and is more robust to ties. Both metrics range from -1 to 1, with values closer to 1 indicating strong alignment in model rankings. To ensure fairness, we adjusted all scores computed with Relaxed Accuracy by subtracting 5%, before comparison. The results are shown in Figure 3.

Each dataset varies in how much it emphasizes visual recognition versus reasoning: *(i)* ChartQA $\rightarrow$ Recognition-heavy, *(ii)* ReachQA $\rightarrow$ Balanced, *(iii)* MATH-Vision $\rightarrow$ Reasoning-heavy

Interestingly, Visual-TableQA rankings align most closely with ReachQA, but not with ChartQA or MATH-Vision individually. This suggests that Visual-TableQA does not favor models that excel solely at recognition or solely at reasoning. Instead, it rewards models capable of both—making it a comprehensive benchmark for evaluating all aspects of visual reasoning.

Visual-TableQA Effectively Transfers to Other Benchmarks: To assess the transferability of Visual-TableQA, we investigated how fine-tuning on Visual-TableQA impacts performance across other benchmarks. As shown in Table 2, supervision from Visual-TableQA led to significant generalization beyond its native domain. Notably, it improved the accuracy of `Qwen2.5-VL-7B-Instruct` on *ReachQA* from 49.23% to 60.95%, and on *MATH-Vision* from 25.10% to 49.77%, despite these datasets not being explicitly table-focused. This finding is further supported by Table 4, which reports similar gains in generalization across three additional models: `LLaVA-Next-Llama3-8B`, `MiniCPM-V2.5-Llama3`, and `InternVL2-8B`.

However, this transferability is not reciprocal. Fine-tuning `Qwen2.5-VL-7B-Instruct` on *ReachQA* alone yields only modest in-domain gains ($49.23\% \rightarrow 55.75\%$) and leads to reduced performance on both *ChartQA* and *Visual-TableQA*. This suggests that Visual-TableQA provides a more generalizable reasoning signal—rooted in layout understanding, symbolic interpretation, and multi-step reasoning—compared to standard benchmarks.

Proprietary Models Outperform Open-Source Models on Average: `Claude 3.5 Sonnet` achieves the highest performance across nearly all benchmarks. However, fine-tuning on Visual-TableQA substantially narrows the gap between proprietary and open-source models. Notably, the performance of `Qwen2.5-VL-7B-Instruct` increases significantly across all evaluated benchmarks—surpassing several state-of-the-art proprietary models, including `GPT-4o`, `GPT-4o-mini`, and `Gemini 2.5 Pro`.

4 Discussion

4.1 Visual-TableQA vs ReachQA

Table 4: Performance of fine-tuned models on the two splits of the ReachQA test set: *Recognition* (Reco) and *Reasoning* (Reas), each consisting of exactly 1,000 samples. Best performances per model category are in **bold**. The values in blue are from our own evaluation using the LLM jury, while the remaining values are taken from He et al..

Model	Reco	Reas	Model	Reco	Reas
LLaVA-Next-Llama3-8B	17.9	6.5	InternVL2-8B	33.7	16.2
+ ReachQA	**29.6**	11.1	**+ ReachQA**	**49.8**	21.3
+ Visual-TableQA	28.4	**20.2**	**+ Visual-TableQA**	45.6	**34.5**
MiniCPM-V2.5-Llama3	25.3	10.3	Qwen2.5-VL-7B-Instruct	66.20	33.10
+ ReachQA	35.10	11	**+ ReachQA**	69.6	40.30
+ Visual-TableQA	**36.20**	**31.50**	**+ Visual-TableQA**	**70.3**	**50.6**
Average gains					
+ ReachQA	$+10.25$	$+4.4$			
+ Visual-TableQA	$+9.35$	$+17.68$			

The ReachQA dataset is divided into two equally sized subsets: *Recognition*, which tests a model's ability to extract relevant information from charts, and *Reasoning*, which evaluates a model's capacity to understand complex and abstract data structures. Table 4 reports the performance gains of multiple fine-tuned models on these two tasks.

On average, models fine-tuned on ReachQA exhibit an accuracy improvement of $+10.25$ points on the *Recognition* task and $+4.4$ points on the *Reasoning* task. In comparison, models fine-tuned on Visual-TableQA show an average gain of $+9.35$ on *Recognition*—a comparable result—but a significantly larger gain of $+17.68$ on *Reasoning*.

This stark contrast in reasoning performance can be attributed to the presence of high-quality rationales in Visual-TableQA annotations, along with the inclusion of more complex and diverse visual structures. In other words, despite being roughly three times smaller than ReachQA in terms of sample count, Visual-TableQA places a stronger emphasis on qualitative richness over quantity. As a result, it appears to enable more effective knowledge distillation, particularly for tasks requiring symbolic interpretation and multi-step reasoning.

4.2 Visual-TableQA's Advantages Compared to Other Datasets

Table 1 shows that only a few table-focused QA datasets—namely TAT-DQA, Table-VQA, and TableVQA-Bench—represent tables as rendered images. **Visual-TableQA** surpasses these by offering richer layout diversity, broader topic coverage, systematic visual complexity, and high-quality rationales. These attributes make it particularly effective for training models with transferable reasoning skills. Supporting this, models fine-tuned solely on **Visual-TableQA**—such as `LLaVA-Next-Llama3-8B`—demonstrated significant gains on external benchmarks (Table-VQA and TableVQA-Bench), as seen in Table 5.

Interestingly, `Qwen2.5-VL-7B-Instruct` did not follow the same performance trend: it showed degradation on tasks such as *VTabFact* (Yes/No fact verification), *VWTQ* (Wikipedia table retrieval), and *VWTQ-Syn* (synthetic variants). To understand this, we manually analyzed its errors before and after fine-tuning on *VTabFact*, categorizing them into eight types: *partial data extraction, hallucination, incoherence, misunderstanding, reasoning errors, evaluation mistakes, dataset ambiguity*, and *annotation flaws*. Results (Figure 14) show that while the total number of errors slightly increased post-finetuning, most now fall into the *incoherence* class, with all other error types significantly reduced. This suggests a sharpening of reasoning patterns but also highlights a need for future work targeting specific error types through synthetic supervision. Further details are provided in Section I.

Beyond transferability and diversity, a key advantage of Visual-TableQA lies in its modularity and scalability as explained in Section 4.3 .

Table 5: Performance of fine-tuned models on Table-VQA test set and the four splits of the TableVQA-Bench dataset: *FinTabNetQA* (finance-related tables), *VTabFact* (table-based fact verification with Yes/No questions), *VWTQ* (information retrieval from Wikipedia tables), and *VWTQ-Syn* (synthetic visual variants of VWTQ). Best performances per model variants are shown in **bold**. Values in blue are from our own evaluation using the `DeepSeek-Prover-v2`, while remaining values are reported from Fu et al. (2025).

Model	TableVQA-Bench				Table-VQA
	FinTabNetQA	VTabFact	VWTQ	VWTQ-Syn	
GPT-4o	96.8	78.0	72.8	82.4	–
LLaVA-Next-34B	–	71.2	36.4	38.0	–
LLaVA-Next-Llama3-8B	52.4	37.2	21.5	24.8	25.84
+ Visual-TableQA	**56.8**	**52.0**	**33.2**	**33.6**	**28.89**
Qwen2.5-VL-7B-Instruct	96.4	**82.0**	**68.53**	**74.0**	**79.03**
+ Visual-TableQA	**97.2**	70.6	61.5	69.6	75.23

4.3 Scalability of the Pipeline and Its Benefits for Knowledge Distillation

This modular pipeline supports scalable generation with a clean separation of concerns—table structure synthesis, QA creation, and validation—making each component independently reusable and upgradable. By automating the entire process from table generation to jury-based quality control, Visual-TableQA provides a cost-efficient and high-quality benchmark for advancing multimodal reasoning over complex visual inputs. A central component of our pipeline is the mechanism of **cross-model inspiration** 2.2, a collaborative prompting strategy. In this process, stronger models generate layout "seeds" that guide weaker models in synthesizing structurally diverse tables, fostering novel visual configurations through iterative transfer. The same principle extends to question–answer generation: models are prompted with both layout and topical cues—often proposed by stronger models—to create new QA pairs. This enables weaker models to contribute meaningfully to the dataset by expanding the range of questions and reasoning patterns. Through this dual-inspiration process, the pipeline cultivates a collaborative multi-model co-creation space, where models of varying capabilities distill collective knowledge not through imitation, but through generative inspiration, while maintaining data quality. In this regard, **Visual-TableQA** distinguishes itself from other synthetic datasets Aboutalebi et al. (2024); Wang et al. (2024a); Li et al. (2025); He et al..

5 Conclusion

In this work, we introduced Visual-TableQA, a large-scale, open-domain, multimodal dataset designed to rigorously evaluate visual reasoning capabilities over complex table images. Building on the principles of Code-as-Intermediary Translation (CIT), we developed a fully automated, modular pipeline for generating LaTeX-rendered tables, reasoning-intensive question–answer pairs, and high-quality rationales—all verified by a jury of strong LLMs. Despite being cost-efficient (generated for under \$100), Visual-TableQA offers unprecedented diversity in table structures, visual features, and reasoning depth. We showed that Visual-TableQA not only challenges existing visual language models (VLMs) but also serves as an effective training signal for improving reasoning performance. Fine-tuning on Visual-TableQA led to substantial gains across multiple benchmarks—both table-centric and general-purpose—including ReachQA and MATH-Vision, demonstrating the dataset's capacity to bridge the performance gap between open-source and proprietary models.

References

Hossein Aboutalebi, Hwanjun Song, Yusheng Xie, Arshit Gupta, Justin Sun, Hang Su, Igor Shalyminov, Nikolaos Pappas, Siffi Singh, and Saab Mansour. Magid: An automated pipeline for generating synthetic multi-modal datasets. *arXiv preprint arXiv:2403.03194*, 2024.

Pranav Agarwal and Ioana Ciucă. Supernova event dataset: Interpreting large language model's personality through critical event analysis. *arXiv preprint arXiv:2506.12189*, 2025.

Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. `https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf`, 2024. Accessed: 2025-08-01.

Anthropic. Claude opus 4 & claude sonnet 4 — system card. `https://www.anthropic.com/claude-4-system-card`, May 2025. Accessed: 2025-08-01.

Shaohan Chen, Yujia Zhang, Xiangpeng Cao, Shaolei He, Chen Zhao, Zhihua Liu, Chongming Li, Jing Liu, Qiang Liu, Fan Liu, et al. Qwen-vl: A versatile vision-language model with image, text, and box comprehension. *arXiv preprint arXiv:2403.18751*, 2024a. URL `https://arxiv.org/abs/2403.18751`.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=rkeJRhNYDH`.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1026–1036, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.91. URL `https://aclanthology.org/2020.findings-emnlp.91/`.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: 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*, pp. 24185–24198, 2024b.

DeepSeek-AI. DeepSeek-R1-Distill-Qwen-32B. `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`, 2025. Model card. Accessed: 2025-08-01.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Google. Gemini 2.0 flash: Model card. `https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf`, 2025a. Published: 2025-04-15. Accessed: 2025-08-01.

Google. Gemini 2.5 flash: Model card. `https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash.pdf`, 2025b. Updated: 2025-06-26. Accessed: 2025-08-01.

Google. Gemini 2.5 pro: Model card. `https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf`, 2025c. Model card. Last updated: 2025-06-27. Accessed: 2025-08-01.

Wei He, Zhiheng Xi, Wanxu Zhao, Xiaoran Fan, Yiwen Ding, Zifei Shan, Tao Gui, Qi Zhang, and Xuanjing Huang. Distill visual chart reasoning ability from llms to mllms.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sahil Kale and Vijaykant Nadadur. Texpert: A multi-level benchmark for evaluating latex code generation by llms. *arXiv preprint arXiv:2506.16990*, 2025.

Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. AIT-QA: Question answering dataset over complex tables in the airline industry. In Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pp. 305–314, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-industry.34. URL `https://aclanthology.org/2022.naacl-industry.34/`.

Maurice George Kendall. Rank correlation methods. 1948.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024.

Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

Andrew Li, Rahul Thapa, Rahul Chalamala, Qingyang Wu, Kezhen Chen, and James Zou. Smir: Efficient synthetic data pipeline to improve multi-image reasoning. *arXiv preprint arXiv:2501.03675*, 2025.

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

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*, 2020.

Meta AI. Llama 4 Maverick 17B-128E Instruct. `https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct`, 2025. Model card. Accessed: 2025-08-01.

Mistral AI. Mistral Small 3.1 24B Instruct. `https://huggingface.co/mistralai/MistralâĂŠSmallâĂŠ3.1âĂŠ24BâĂŠInstructâĂŠ2503`, 2025. Model card. Accessed: 2025-08-01.

OpenAI. GPT-4o. `https://openai.com/index/gpt-4o`, 2024. Accessed: 2025-07-30.

OpenAI. GPT-4.1. `https://openai.com/index/gpt-4-1/`, 2025a. Accessed: 2025-08-01.

OpenAI. GPT-4o-mini. `https://platform.openai.com/docs/models/gpt-4o`, 2025b. Accessed: 2025-08-01.

OpenAI. OpenAI o3 Reasoning Model. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025. Accessed: 2025-07-31.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1142. URL `https://aclanthology.org/P15-1142`.

Qwen Team. Qwen3-30B-A3B. `https://huggingface.co/Qwen/Qwen3-30B-A3B`, 2025a. Model card. Accessed: 2025-08-01.

Qwen Team. Qwen3-QwQ-32B. `https://huggingface.co/Qwen/QwQ-32B`, 2025b. Model card. Accessed: 2025-08-01.

Reka AI. Reka Flash 3. `https://huggingface.co/RekaAI/reka-flash-3`, 2025. Model card. Accessed: 2025-08-01.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ee6W5UgQLa.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

TNG Technology Consulting GmbH. Deepseek-r1t-chimera, April 2025. URL https://huggingface.co/tngtech/DeepSeek-R1T-Chimera.

Cyrile Delestre Tom Agonnoude, 2024. URL https://huggingface.co/datasets/cmarkea/table-vqa.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Jiankang Wang, Jianjun Xu, Xiaorui Wang, Yuxin Wang, Mengting Xing, Shancheng Fang, Zhineng Chen, Hongtao Xie, and Yongdong Zhang. A graph-based synthetic data pipeline for scaling high-quality reasoning instructions. *CoRR*, 2024a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=QWTCcxMpPA.

xAI. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3, 2025. Accessed: 2025-08-01.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 697–706. IEEE, 2021.

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

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4857–4866, 2022.

Appendix

A Extended Related Works

The vast majority of table-based QA datasets—such as HybridQA Chen et al. (2020b), WikiTableQuestions Pasupat & Liang (2015), WikiSQL Zhong et al. (2017), and AIT-QA Katsis et al. (2022)—represent tables in textual format rather than as rendered images, thereby bypassing the challenges associated with visual layout interpretation. In contrast, our work focuses exclusively on multimodal datasets—those that contain both textual and visual (image-based) information. These can generally be grouped into two main categories: real-world datasets, collected from authentic documents, and synthetic datasets, generated using automated tools. Real-world multimodal QA datasets that emphasize tables—such as TAT-DQA Zhu et al. (2022), and TableVQA-Bench Kim et al. (2024)—tend to be highly domain-specific, limiting diversity in both table layouts and question types. For example, TAT-DQA Zhu et al. (2022) combines tabular and textual data from financial reports and, while it introduces hybrid contexts for realistic reasoning, its questions rely heavily on reading textual input rather than interpreting visual structure. Similarly, TableVQA-Bench Kim et al. (2024) consists of 83% real-world tables (1,250 out of 1,500), primarily sourced from task-specific datasets such as WikiTableQuestions (information retrieval), TabFact Chen et al. (2020a) (fact verification), and FinTabNet Zheng et al. (2021) (financial data extraction). Due to its relatively small size and the specialized nature of its subsets, the dataset exhibits limited visual diversity. This limitation also extends to the remaining 16% synthetic tables, whose visual variation is restricted to basic formatting attributes such as background color, border size, font size, and style. More recently, ChartQA Masry et al. and DocVQA Mathew et al. (2020) have introduced large open-domain datasets for visual question answering. ChartQA focuses on reasoning over charts and plots; however, its tasks primarily involve shallow reasoning and do not reflect the structural complexity and layout diversity found in real-world tables. In contrast, DocVQA offers greater diversity in document layouts and structures, but lacks significant visual challenge—recent VLMs, including relatively lightweight models like `Qwen2.5-VL-7B-Instruct` Team (2025), achieve over 94% accuracy on this benchmark.

Among synthetic datasets, *MultiModalQA* Talmor et al. (2021) stands out as the only open-domain resource focused on tables. It combines real-world figures, diagrams, and text passages sourced from Wikipedia, with QA pairs crafted to assess both reasoning and visual comprehension. Although it incorporates real content, the dataset is considered synthetic due to the way it links independent modalities and generates QA pairs through formalized templates. However, this approach results in limited diversity, as the questions are derived from a finite set of templates. In contrast, *Table-VQA* Tom Agonnoude (2024) is a fully synthetic dataset generated using state-of-the-art LLMs. Nonetheless, it lacks visual diversity and complexity—its tables follow similar formats and are predominantly centered around technical domains such as statistics, physics, and algorithms, all of which are heavily numerical in nature.

Most of real-world datasets rely heavily on manual labeling, data collection, and preprocessing—factors that significantly constrain their scalability. Recently, ReachQA He et al. introduced a more scalable and innovative approach through its Code-as-Intermediary Translation (CIT) pipeline. This method generates synthetic charts and reasoning questions by leveraging textual intermediaries such as Python code, demonstrating that advanced reasoning capabilities of large language models (LLMs) can be effectively transferred to visual models. While ReachQA successfully addresses both scalability and reasoning complexity, its approach is tailored to chart-based visualizations and does not extend to structured tabular data.

To summarize, existing table-based benchmarks consistently fall short in one or more key areas: visual diversity, reasoning depth, or scalability. Notably, aside from MultiModalQA, there is no open-domain dataset designed to evaluate model performance on rendered table images, despite their prevalence in real-world settings such as reports, academic papers, and spreadsheets. In this work, we introduce **Visual-TableQA**, a multimodal open-domain synthetic dataset specifically created to assess reasoning capabilities over table images using LLMs

B Considerations and Limitations

The main limitations of this work relate to the use of Code-as-Intermediary Translation (CIT) He et al. and the assessment of data quality. While we adopted LaTeX as an intermediate representation for tables, its expressiveness is limited when handling more complex or visually rich images. Developing a robust, bidirectional image-to-text encoding system remains an open and promising area for future research. In terms of data quality evaluation, although automatic metrics such as ROSCOE provide useful insights, they are not yet as reliable as human judgment. As a result, human annotators continue to play a critical role in ensuring high-quality data, especially when scaling synthetic datasets for reasoning tasks.

In addition, we also observed that certain models, such as `Qwen2.5-VL-7B-Instruct`, did not consistently benefit from Visual-TableQA supervision across all downstream tasks, highlighting a potential limitation in generalization that warrants further investigation.

C TableQA Layout and Topic Diversity

As described in Section 2.2, we sampled 5,000 distinct topics using `GPT-4o` to serve as inspirations for table generation. To better illustrate topic diversity, we grouped these topics into 20 semantic clusters using the K-Means algorithm. Figure 5 displays a 2D projection of these clusters, where each color represents a distinct semantic group. For the 12 largest clusters, we highlight representative topics to give a sense of their thematic content.

In addition, Figure 4 shows the cumulative percentage of topics covered as clusters are added in descending order of size. The smooth progression of the curve indicates that the clusters are relatively uniform in size, confirming a balanced distribution of topic diversity throughout the dataset.

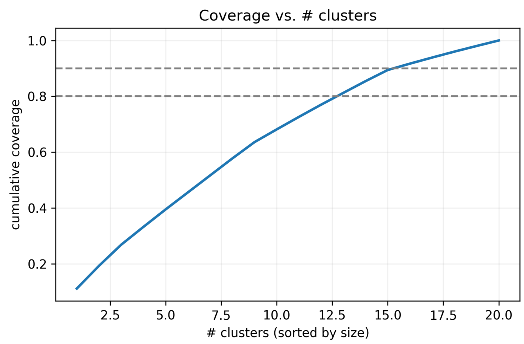
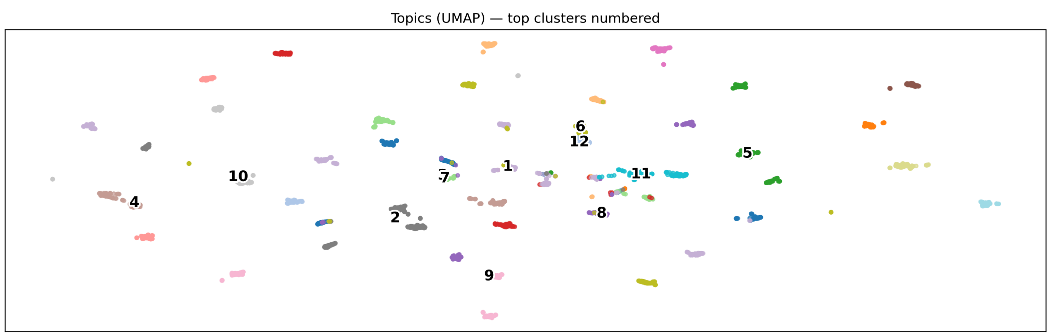


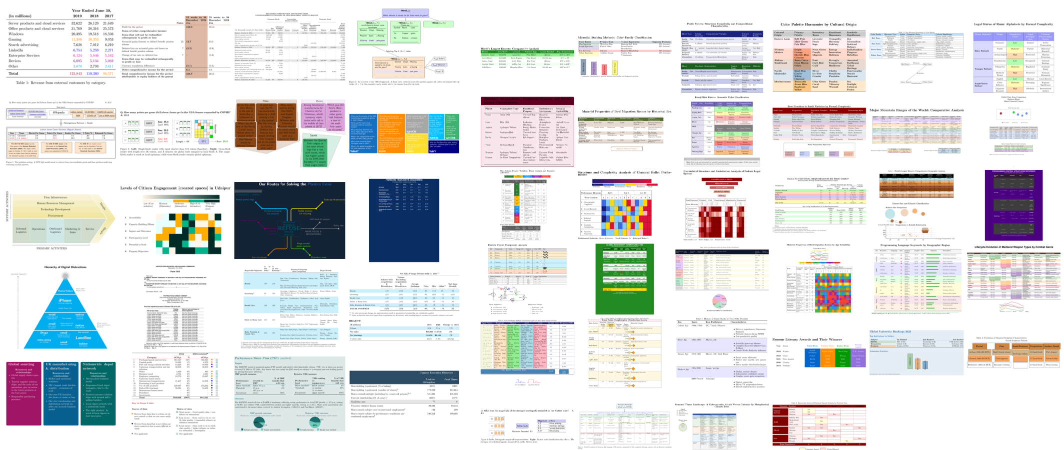
Figure 4: Cumulative topic coverage as clusters are added by descending size. The uniform slope indicates an even distribution of topics across clusters.



#	Cluster (alias)	Count	#	Cluster (alias)	Count
1	Festival calendar of poetic meters by material / Festival calendar of world mythologies by genre / Timeline of poetic meters by cultural origin	556	7	Catalogue of linguistic case systems by material / Catalogue of linguistic case systems by cultural origin / Catalogue of linguistic case systems by	304
2	Glossary of vintage camera formats by construction method / Morphology of vintage camera formats by genre / Genealogy of vintage camera formats by	408	8	Catalogue of cloud formations by symbolic meaning / Symbolism in cloud formations by climate zone / Catalogue of cloud formations by genre	304
3	Symbolism in famous algorithmic paradigms by common usage / Symbolism in programming language keywords by formal complexity / Catalogue of	378	9	Taxonomy of breathing techniques in meditation by era / Catalogue of breathing techniques in meditation by genre / Taxonomy of yoga poses by	293
4	Classification of tea ceremony utensils by flavor profile / Classification of tea ceremony utensils by geographic region / Taxonomy of tea ceremony	319	10	Checklist of heraldic tinctures by genre / Catalogue of heraldic tinctures by material / Catalogue of heraldic tinctures by functional	228
5	Genealogy of traditional weaving patterns by learning curve / Catalogue of traditional weaving patterns by functional purpose / Taxonomy of knot	313	11	Genealogy of chess openings by functional purpose / Taxonomy of chess openings by era / Lifecycle stages of chess openings by genre	224
6	Taxonomy of sushi varieties by genre / Catalogue of sushi varieties by target audience / Classification of sushi varieties by flavor	304	12	Symbolism in plant leaf shapes by cultural origin / Catalogue of plant leaf shapes by symbolic meaning / Catalogue of flower language meanings by	219

Figure 5: 2D projection of the 5,000 topics using UMAP and K-Means clustering. Each color denotes a semantic cluster. Representative topics are listed for the 12 largest clusters to illustrate diversity.

To illustrate the diversity of table layouts produced by our pipeline, Figure 6 displays a side-by-side comparison between the initial seed tables used in the first generation iteration and a sample of layouts generated in subsequent steps. The wide range of structures highlights the pipeline's capacity to create rich and varied visual designs from limited starting templates.



(a) Layout seeds used for the initial table generation.

(b) Sample of diverse layouts generated by our pipeline.

Figure 6: Visual diversity of table layouts. Left: seed layouts used during the first iteration of table generation. Right: layouts generated through cross-model inspiration and iterative refinement.

D Table Generation Settings

The first stage of the generation pipeline involves **LLM-1**, which is responsible for producing new tables based on given inspirations. Specifically, it receives one LaTeX-formatted table as a *layout inspiration* and three distinct *topic inspirations*. Based on these, it generates three new tables, each aligned with one of the provided topics while drawing structural influence from the layout example. The full prompt used to guide LLM-1 during this step is shown in Figure 7.

The second stage of the generation pipeline involves **LLM-2**, which is responsible for producing question–answer (QA) pairs based on a single LaTeX table. The full prompt used to guide LLM-2 is shown in Figure 8.

To encourage creativity while maintaining factual accuracy, the temperature parameter for each model during this phase is set to 0.7.

In Table 6, we present the average validity rates of the generated QA pairs for each model involved in the QA generation phase.

Table 6: Average QA pair validity across different QA generation models. Accuracies are computed from a sample of at least 500 QA pairs per model.

Model	Acc. (%)
Llama 4 Maverick 17B-128E Instruct	88
Gemini 2.0 Flash	89.1
Gemini 2.5 Flash	93.1
Gemini 2.5 Pro	89.3
GPT-4.1	90.4
Qwen3-30B-A3B	76.6
Qwen3-QwQ-32B	92.6
DeepSeek-R1-Distill-Qwen-32B	73.4
DeepSeek-R1T-Chimera	89.4
Claude Sonnet 4	91
Claude 3.5 Haiku	90.2
Grok 3 Beta	89.4
Reka Flash 3	79

E Common Anomalies in Generated Tables

As summarized in Table 3, several tables generated by different LLMs required post-processing adjustments to ensure visual clarity, LaTeX correctness, and topic alignment. Below, we categorize the most common types of anomalies observed during generation:

```
You are an expert in generating synthetic datasets composed of LaTeX-formatted tables, optionally
      accompanied by illustrative diagrams. Your task is to produce structured content suitable for data-
      centric documents, ensuring each table (and diagram, if included) is clear, well-organized, and
      visually informative.
Your final output should start with ```json and end with ``` as plain text, not just formatting. Like this:


```json
{
  "table_1": "BEGIN_LATEX
<LaTeX code for table 1 (with/without diagram) here>
END_LATEX",
  "table_2": "BEGIN_LATEX
<LaTeX code for table 2 (with/without diagram) here>
END_LATEX",
  "table_3": "BEGIN_LATEX
<LaTeX code for table 3 (with/without diagram) here>
END_LATEX"
}
```

Requirements:
    The tables and diagrams will be used to generate reasoning questions. Therefore:

        - If topic inspirations are supplied, ensure every generated table aligns with those topics.
        - Each LaTeX output must primarily consist of a table. Include a diagram only if it meaningfully
              complements the table; avoid adding one unnecessarily. Do not generate diagrams alone. If a
              diagram is empty or non necessary DON'T INCLUDE it.
        - Keep any diagram minimal--smaller than the table, chart-free, and purely illustrative--serving
              only to reinforce the table's content without adding new information.
        - Each table and their diagram must contain realistic, domain-relevant content. They must be self-
              contained, include a clear descriptive title and not rely on external data to compile.
        - The type of information presented should be diverse--such as numerical data or qualitative. The
              variety and richness of visual elements is essential to the overall quality of the table and
              their diagram. Table quality should also come with a large number of rows and columns.
        - Table and diagram layouts should be creatively designed--taking inspiration from reference
              example (when provided) but incorporating meaningful variations such as colors, multi-row or
              multi-column cells, custom formatting adjustments, or any other visual enhancement that
              promotes structural diversity.
        - Table layouts should be at least as complex as the example provided, don't try to simplify (
              diagrams are not mandatory). Table complexity should also come with a large number of rows and
               columns.
        - Do NOT escape any characters in the LaTeX code. The LaTeX must be written as plain text, exactly
              as it would appear in a .tex file, with real line breaks and single backslashes (\), not JSON-
              escaped.
        - All LaTeX tables and diagrams must be constrained to fit entirely within the printable area of a
              standard A4 page when compiled to PDF, without overflowing horizontally or vertically. Use
              appropriate formatting techniques such as adjusting column widths, reducing font size, or
              enabling landscape mode if necessary but NEVER rotation.
        - Make sure each LaTeX table and diagram includes all required \usepackage declarations and is
              enclosed within a complete, compilable LaTeX document structure, including the appropriate
              preamble and \begin{document}...\end{document} block.
        - Make sure each LaTeX codes start and end with BEGIN_LATEX and END_LATEX, respectively.
        - Make sure to wrapp your final answer with ```json at the beginning and ``` at the end.
```

Figure 7: LLM prompt used for table generation.

Layout and Formatting Issues

- **Table overflow:** Tables exceeding page margins due to improper column widths or missing column dimensioning.
- **Text overflow:** Cell content spilling outside the cell boundary, especially in narrow columns or with long strings.
- **Invisible content:** Multirow cells with background colors that obscure cell text (e.g., white text on white background).
- **Improper horizontal lines:** `\midrule` or `\hline` splitting across multirow cells, breaking visual coherence.

Content Relevance and Correctness

- **Empty or irrelevant tables:** Tables with placeholder content or unrelated to the assigned topic.

```
You are an expert in generating question-answer pairs from LaTeX-formatted. Your task is to create a
     structured dataset consisting of visually challenging, reasoning-based questions and their
     corresponding answers derived from a given LaTeX formatted table with optional diagram.

Input:

You will be provided with a sample LaTeX table as context. Based on this table or diagram, your goal is to
      generate a JSON object with the following structure:

   questions: A python list of 3 challenging questions that require reasoning and analysis based ONLY on
        the data presented in the table and the optional diagram. The questions must be answerable using
        ONLY the information in the table or diagram(no extra knowledge).
   answers: A python list of 3 detailed answers to the 3 questions, including a clear chain of thought
        explaining the reasoning process.

Requirements:

   All questions must be relevant to the table's context and designed to test deeper understanding or
        inference.
   When possible, all questions should make full use of the visual or structural elements of the table or
        diagram (such as rows, columns, headers, colors, patterns, diagrams etc.) while maintaining clear
         relevance to the table’s content.
   Questions must be clear and answerable with an objective methodology, no subjective question.
   All entries (both questions and answers) should be returned as lists of string values.
   The global result should be a single JSON object wrapped in a markdown code block using ```json at the
        beginning and ``` at the end, and containing all two key-value pairs.
   This means your output should start with ```json and end with ``` as plain text, not just formatting.
```

Figure 8: LLM prompt used for QA generation.

- **Incorrect topic alignment:** Generated tables that do not match the intended topic inspiration.
- **Duplicate outputs:** All three tables generated for a prompt are identical or nearly identical in structure/content.
- **Missing external resources:** References to images or figures not included or available in the output.
- **Incorrect LaTeX syntax:** Math symbols placed outside of math environments, leading to compilation errors.

Diagram-Specific Issues

- **Missing tables:** Some generations return only a diagram without an accompanying table.
- **Node placement errors:** Overlapping or misaligned nodes in TikZ diagrams.
- **Arrow misplacement:** Arrows that do not connect to correct nodes or that overlap diagram elements improperly.
- **Legend/title confusion:** Titles or legends positioned incorrectly or detached from the relevant diagram elements.
- **Visual inconsistencies:** General drawing flaws, such as missing anchors, inconsistent line styles, or unintended overlaps.

These issues highlight the need for a validation loop in the TableQA pipeline and justify the inclusion of human verification stages to ensure dataset quality.

F ROSCOE Metric Scores

In addition to the LLM jury validation of the Visual-TableQA dataset, we conducted a complementary quality assessment using the ROSCOE framework Golovneva et al.. This evaluation measures step-by-step reasoning coherence across multiple dimensions, including semantic alignment, logical consistency, and contextual grounding. The resulting scores, reported in Table 7, further support the reliability and high quality of the generated tables and QA pairs, reinforcing the effectiveness of our data generation pipeline.

Table 7: ROSCOE Golovneva et al. reasoning metrics averaged over the whole dataset. The "Direction" column indicates whether higher or lower values correspond to better performance for each metric.

Metric	Direction	Mean	Std
Semantic Adequacy (↑)			
Faithfulness-Step	↑	0.99	5e-4
Informativeness-Step	↑	0.99	5e-3
Informativeness-Chain	↑	0.98	1.4e-2
Faithfulness-Token	↑	0.99	2e-3
Avg		**0.99**	
Redundancy & Risk (↓)			
Repetition-Token	↓	0.06	0.12
Repetition-Step	↓	0.06	0.12
Avg		**0.06**	
Logical Inference (↑)			
Discourse-Representation	↑	0.68	0.41
Coherence-Step	↑	0.7	0.40
Avg		**0.69**	
Fluency & Perplexity (↓)			
Perplexity-Step	↓	0.01	0.01
Perplexity-Chain	↓	0.05	0.03
Perplexity-Step-Max	↓	8e-3	8e-3
Avg		**0.02**	
Grammaticality (↑)			
Grammar-Step	↑	0.96	0.05
Grammar-Step-Max	↑	0.9	0.13
Avg		**0.93**	

G LLM Jury Reliability

Following recent studies He et al.; Fu et al. (2025); Verga et al. (2024), we adopt a high-performing LLM jury combined with a majority-vote strategy to evaluate model predictions. This multi-model jury setup enhances evaluation robustness and mitigates individual model biases. The configuration details for all jury models are provided in Section G.1. As argued in Verga et al. (2024), aggregating judgments from several strong LLMs yields more consistent and reliable evaluations than relying on a single model. In Section G.2, we discuss the challenges of LLM-based evaluations and we report a detailed comparison between our LLM jury and human annotators on `Qwen2.5-VL-7B` predictions across multiple benchmarks. In Section G.3, we provide a detailed analysis of jury-to-jury agreement across both table and QA pair quality assessments, as well as benchmark evaluations. The results reveal strong inter-annotator alignment, validating the consistency and effectiveness of our jury-based evaluation protocol.

G.1 LLM Jury Settings

LLM juries were involved at two key stages of the pipeline: (i) **quality assurance**, where generated tables and QA pairs were validated before inclusion in the dataset, and (ii) **evaluation benchmarking**, where model responses were assessed for accuracy and reasoning quality.

The prompt used for quality assurance is shown in Figure 9, while the prompt used for evaluation during benchmarking is shown in Figure 10. Each jury consisted of multiple high-performing vision-language or reasoning-capable LLMs, and final decisions were made via majority voting.

To ensure consistency and reproducibility across evaluations, all LLM jury calls were executed with a temperature setting between 0.0 and 0.1.

We observed a notable drop in judgment accuracy when LLM juries were instructed to return only a structured JSON verdict without any preceding explanation. In particular, `Mistral-large` systematically omitted its reasoning whenever the keyword `JSON` was included in the prompt. This issue was effectively mitigated by explicitly instructing models to provide a rationale prior to their final decision and by avoiding any direct mention of `JSON` in the prompt, except for `GPT-4.1` which remained robust under such formatting. Including explicit reasoning significantly improved the reliability and depth of model evaluations, likely by reducing shallow assessments and prompting more thoughtful judgments.

```
You are a reasoning question answer expert. You will be given a LaTeX formated table with/without diagram,
      a list of 3 topics, and a pair of a question and its answer.

Your task is to evaluate the pair of question answer based solely on the data in the LaTeX code and these
     criteria:

   1) Does the LaTeX code contain a Table (not some charts alone or diagrams alone) ?
   2) Does the table, any optional diagrams, and the rest of the document are on one single topic from the
        provided list of topics, and internally consistent (be careful to off-topic diagrams)?
   3) Is the question clear and related to the table or the diagram?
   4) Is the answer (including its reasoning) totally valid and does it actually respond to the question?
   5) Is the answer FULLY supported by and ONLY BY the table or diagram data (no extra knowledge)?

If the five criteria are true, mark the pair as correct.
If one of the criteria is not met, mark it as incorrect.

Think step by step and conclude with your decision and the index of the criterium not met (if none, index
     is 0) as follows:
JSON_mention
{{"decision": [0, index_of_the_criterium_not_met]}} for incorrect or {{"decision": [1, 0]}} for correct.
```

Figure 9: LLM prompt used for QA evaluation.

```
You are an expert evaluator of question-answer pairs. You will be given a question a model's answer and a
     ground truth answer (reference).
Evaluate the answer based on these criteria:
1) Is the model's answer logically consistent?
2) Does the model's answer convey the same meaning as the ground truth?
If the two criteria are true, mark the pair as correct. If one of the criteria is not met, mark it as
     incorrect.

Think step by step and conclude with your verdict and the index of the criterium not met (if none, index
     is 0) as follows:
{{"verdict": [0, index_of_the_criterium_not_met]}} for incorrect or {{"verdict": [1, 0]}} for correct.

Question: {question}
Answer: {answer}
Ground Truth: {ground_truth}
Response:
```

Figure 10: LLM prompt used for Benchmark evaluation.

G.2 LLM Jury Limitations and Mitigation Strategies

Our evaluation protocol measures alignment between a model's prediction and the reference answer, leveraging LLM juries for semantic comparison. This strategy is highly cost-effective, enabling scalable automatic assessments on reasoning tasks. However, it comes with important limitations. A model's response may rely on external world knowledge rather than explicitly extracting information from the table image. In such cases, LLM juries lack the ability to verify whether the response is grounded in the table itself, as they do not have access to the visual context. This weakness makes the evaluation pipeline vulnerable to false positives, especially when the model outputs a factually correct answer that is not actually derivable from the table content. This could be mitigated by including rendered table images alongside the question and model prediction at evaluation time—giving LLM juries full visual grounding. Unfortunately, While effective, this approach is significantly more costly in terms of API calls and inference latency, making it challenging to scale on large datasets. These observations underscore the need for hybrid evaluation strategies combining automatic LLM-based judgments with human verification, in order to quantify the risk of jury-related errors at scale.

To rigorously assess this risk, we conducted a manual evaluation of `Qwen2.5-VL-7B-Instruct`'s predictions on both Visual-TableQA and the *VTabFact* split of TableVQA-Bench, one of the benchmarks where the model exhibited degraded performance after fine-tuning. Our manual analysis serves as a ground truth reference to assess the reliability and limitations of LLM-based jury evaluation. Our evaluation considers not only the final answer, but also the validity of the reasoning process: even predictions with the correct answer are marked as incorrect if the chain of thought is flawed. We used the manual analysis to validate the judgments made by the LLM juries. We identified two types LLM Juries errors:

- *False Positive:* The LLM jury mistakenly accepts an incorrect model prediction as correct.
- *False Negative:* The LLM jury mistakenly rejects a correct model prediction as incorrect.

The errors distribution are detailed in Table 8. The results reveal that the LLM jury aligns with human judgment within a reasonable margin of 4.7%, even under our strict annotation protocol that penalizes incorrect reasoning regardless of the final answer. A more relaxed evaluation criterion would yield an even smaller discrepancy. These findings reinforce the reliability of our LLM jury setup, demonstrating its effectiveness as a scalable proxy for human evaluation.

Table 8: Misclassifications by LLM juries on Visual-TableQA and *VTabFact*. Percentages are calculated relative to the total number of evaluated examples in each dataset.

Error Type	**Visual-TableQA**	***VtabFact***
False Positive	3.45%	1.6%
False Negative	1.2%	0.0%
Total	4.65%	1.6%

G.3 LLM Jury Agreement Analysis

Figure 11 presents a detailed breakdown of agreement levels between individual LLM juries, as well as their alignment with majority-vote annotations. This analysis was conducted across both table and QA pair quality assessments. The analysis reveals a spectrum of consistency across juries, with `GPT-4.1` emerging as the most reliable, likely due to its robust handling of edge cases. Among all models, proprietary LLMs such as `Gemini-2.5-pro` and `GPT-4.1` show the strongest alignment with the majority vote, while `Deepseek-v3.1` exhibits the weakest agreement. Notably, pairwise jury agreement patterns appear correlated with the models' reasoning capabilities. Despite variability in alignment strength, all juries demonstrate a meaningful degree of concordance with the majority, underscoring the robustness of our collective evaluation protocol.

Conversely, Figure 12 shows consistently strong jury agreement across all models for benchmark evaluations, with no notable divergence between proprietary and open-source LLMs. This can be attributed to the relatively simpler nature of the task (semantic comparison between model predictions and ground truth) compared to the more complex reasoning required for evaluating table-question-answer triples, as analyzed in Figure 11.

G.4 Evaluator Consistency Compared to the Literature

In line with recent studies He et al.; Fu et al. (2025); Verga et al. (2024); Agarwal & Ciucă (2025), we employed a high-performing LLM jury with a majority-vote strategy to evaluate model predictions. The jury consisted of `Mistral-large`, `Deepseek-v3.1`, `Gemini-2.5-pro`, `GPT-4.1`, and `Deepcogito-v2`. In Table 2, baseline values (in black) are reported from He et al., who used `GPT-4o` as the sole evaluator. While our evaluation pipeline involves a broader and more powerful set of models, making it arguably more reliable and robust (Verga et al., 2024), we still consider the two evaluation protocols broadly comparable. In fact, due to the stricter majority-vote requirement across diverse models, our approach may even yield more demanding or rigorous evaluations. This comparability also holds for Table 5, where baseline performances (in black) are taken from Fu et al. (2025), who employed `GPT-4` for their evaluation. We argue that despite methodological differences, all evaluations are consistent enough to be analyzed jointly for comparative purposes.

H Model Finetuning Hyperparameters

The hyperparameters used for LoRA are reported in Table 9. For `Qwen2.5-VL-7B-Instruct`, we employed a two-phase fine-tuning strategy. In the first phase (Tier A), we adapted the text-side modules together with the multi-modal projector, while keeping the vision tower frozen. In the

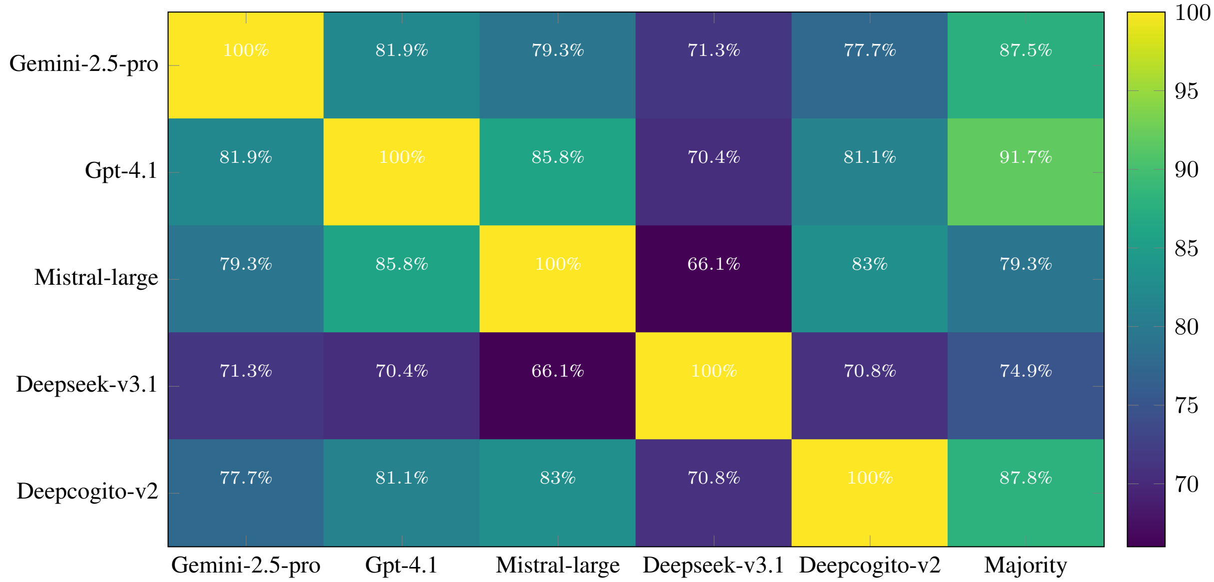


Figure 11: Pairwise agreement (%) between LLM juries, plus alignment with majority vote for tables and QAs quality assessment

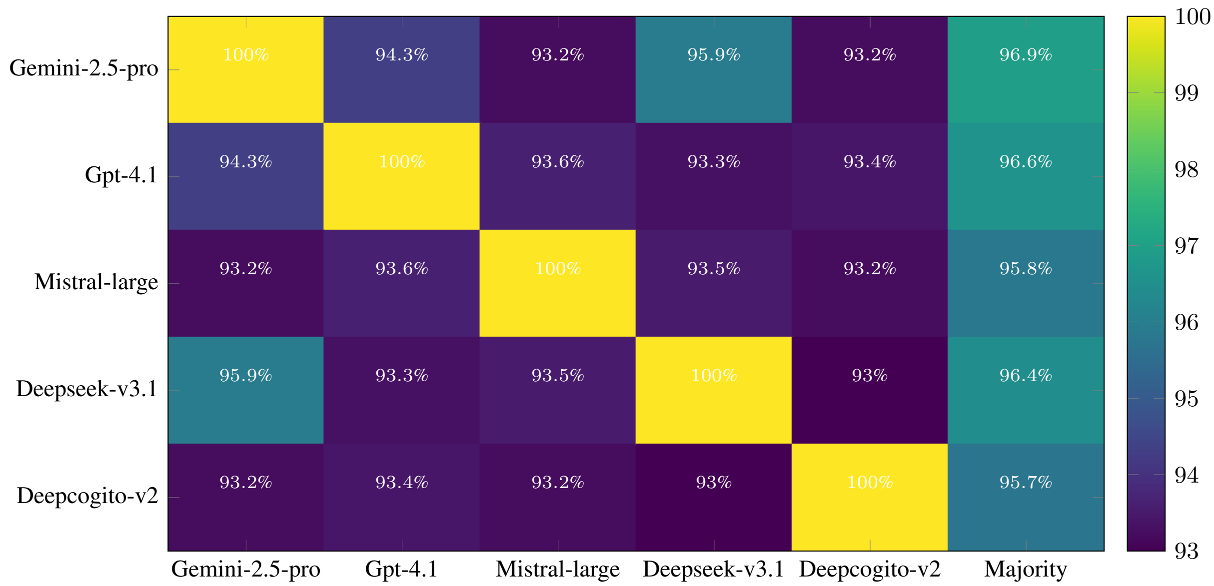


Figure 12: Pairwise agreement (%) between LLM juries, plus alignment with majority vote for benchmark evaluation

second phase (Tier B), we further enabled LoRA on the attention projections of the last four vision blocks, leaving the remaining vision layers untouched.

I Error Taxonomy of Model Predictions

To better understand the failure modes of `Qwen2.5-VL-7B-Instruct`, we conducted a fine-grained manual analysis of its predictions, both before and after fine-tuning on Visual-TableQA and the *VTabFact* split of TableVQA-Bench. This section is organized as follows: *(i)* we define the full set of error types used in our annotation protocol (Section I.1); *(ii)* we present a comparative analysis of errors observed on *VTabFact* (Section I.2).

Table 9: Hyperparameters Used for Fine-Tuning with LoRA. More details in our GitHub repository. Abbreviations: lr=learning rate, r= LoRA rank, α= LoRA α, Targets=targets modules for LoRA.

Model	lr	r	α	Targets
LLaVA-Next-Llama3-8B MiniCPM-V2.5-Llama3	2e-5	16	8	all-linear (llm frozen)
InternVL2-8B		16*	32*	
Qwen2.5-VL-7B-Instruct	1e-4	16	8	Tier A: projector modules**
	2e-5	8	32	Tier B: last 4 vision blocks' attention

* The InternVL training source code sets the LoRA alpha as twice the LoRA rank, as shown in their official implementation here. We followed this convention for full reproducibility and assumed that other baselines applied the same rule.
** We left the Vision Tower untouched as it significantly degraded model performance.

I.1 Error Categories

We classified the observed errors into eight categories (Figure 13):

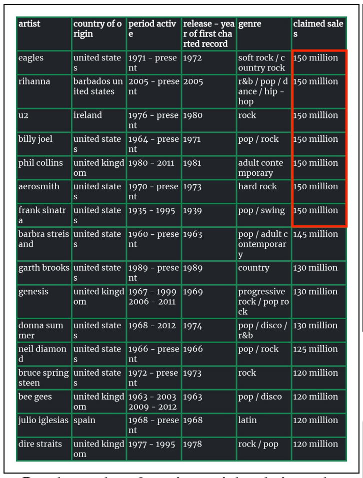
- **Partial Data Extraction:** The model overlooks some relevant entries (e.g., stops counting too early, Figure 13a).
- **Hallucination:** The model references information not present in the table (Figure 13b).
- **Incoherence:** The model extracts the correct data but then misinterprets it or contradicts itself later (deductive error, Figure 13c).
- **Misunderstanding:** The model produces factual statements that do not actually answer the question (Figure 13d).
- **Faulty Methodology/Reasoning:** The reasoning is too shallow, or the model fails to satisfy all the constraints of the query (Figure 13e).
- **Ambiguity (Gray Area):** Both the ground truth and the model's prediction can be reasonably justified (Figure 13f).
- **Dataset Mistake:** The original dataset contains annotation or label errors.

I.2 Error Analysis on VTabFact

The results of our analysis are shown in Figure 14. The side-by-side comparison of Figure 14a and Figure 14b reveals a significant shift in error distribution after fine-tuning: although the total number of errors increases from 67 to 88 (out of 250 samples), most newly introduced errors concentrate in the *Incoherence* and *Hallucination* categories, while all other error types show a marked decline. In particular, *Faulty Methodology/Reasoning* errors are significantly reduced, indicating that the fine-tuned model exhibits more consistent and structured reasoning patterns. This trend is further supported by Figure 14c, which shows the distribution of errors corrected by the fine-tuned model: improvements span across all error categories. In contrast, Figure 14d shows that newly introduced errors after fine-tuning are largely concentrated in only a 2 categories. We attribute the sharp rise in *Incoherence* errors to two key factors:

Degraded Arithmetic Capability: A recurrent issue post-finetuning involves numerical comparison errors or basic math computation errors, leading to faulty deductions despite otherwise correct reasoning steps (see Figure **??**). Such errors were less frequent in the pretrained model.

Answer Template Behavior: The fine-tuned model tends to state its final answer before providing its chain of thought, often declaring an incorrect answer followed by a valid justification that leads to the correct conclusion (see Figure **??**). Despite the correctness of the rationale, we count such instances as errors. In contrast, the pretrained model typically presents its reasoning first, then concludes—leading to fewer coherence violations.

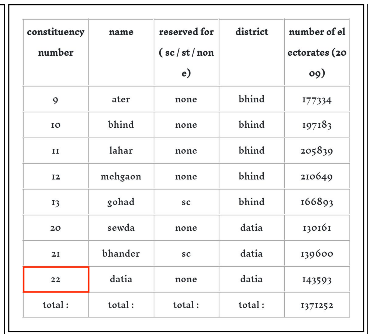


artist	country of origin	period active	release - year of first charted record	genre	claimed sales
eagles	united states	1971 - present	1972	soft rock / country rock	150 million
rihanna	barbados united states	2005 - present	2005	r&b / pop / dance / hip - hop	150 million
u2	ireland	1976 - present	1980	rock	150 million
billy joel	united states	1964 - present	1971	pop / rock	150 million
phil collins	united kingdom	1980 - 2011	1981	adult contemporary	150 million
aerosmith	united states	1970 - present	1973	hard rock	150 million
frank sinatra	united states	1935 - 1995	1939	pop / swing	150 million
barbra streisand	united states	1960 - present	1963	pop / adult contemporary	145 million
garth brooks	united states	1989 - present	1989	country	130 million
genesis	united kingdom	1967 - 1999 2006 - 2011	1969	progressive rock / pop rock	130 million
donna summer	united states	1968 - 2012	1974	pop / disco / r&b	130 million
neil diamond	united states	1966 - present	1966	pop / rock	125 million
bruce springsteen	united states	1972 - present	1973	rock	120 million
bee gees	united kingdom	1963 - 2003 2009 - 2012	1963	pop / disco	120 million
julio iglesias	spain	1968 - present	1968	latin	120 million
dire straits	united kingdom	1977 - 1995	1978	rock / pop	120 million

Q: there be 6 artist with claimed sale of 150 million?

A: To determine if there are 6 artists with claimed sales of 150 million, let's count the entries in the "claimed sales" column where the value is 150 million:
1. Eagles [...] 6. Aerosmith
There are indeed 6 artists listed with claimed sales of 150 million.

(a) Partial Data Extraction

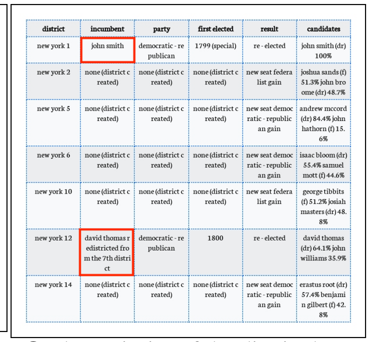


constituency number	name	reserved for (sc / st / none)	district	number of electorates (2009)
9	ater	none	bhind	177334
10	bhind	none	bhind	197183
11	lahar	none	bhind	205839
12	mehgaon	none	bhind	210649
13	gohad	sc	bhind	166893
20	sewda	none	datia	130161
21	bhander	sc	datia	139600
22	datia	none	datia	143593
total :	total :	total :	total :	1371252

Q: datia 's constituency number be lower than gohad 's by 10

A: To determine if Datia's constituency number is lower than Gohad's by 10, we need to compare their respective constituency numbers.
1. Identify the constituency number for Datia: The table shows that Datia's constituency number is 20. [...]

(b) Hallucination

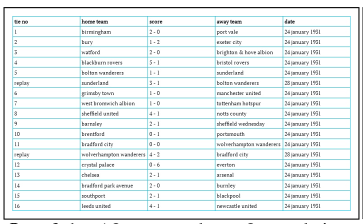


district	incumbent	party	first elected	result	candidates
new york 1	john smith	democratic - republican	1799 (special)	re - elected	john smith (dr) 100%
new york 2	none (district created)	none (district created)	none (district created)	new seat federalist gain	joshua sands (f) 51.3% john broome (dr) 48.7%
new york 5	none (district created)	none (district created)	none (district created)	new seat democratic - republican gain	andrew mccord (dr) 84.4% john hathorn (f) 15.6%
new york 6	none (district created)	none (district created)	none (district created)	new seat democratic - republican gain	isaac bloom (dr) 55.4% samuel mott (f) 44.6%
new york 10	none (district created)	none (district created)	none (district created)	new seat federalist gain	george tibbits (f) 51.2% josiah masters (dr) 48.8%
new york 12	david thomas redistricted from the 7th district	democratic - republican	1800	re - elected	david thomas (dr) 64.1% john williams 35.9%
new york 14	none (district created)	none (district created)	none (district created)	new seat democratic - republican gain	erastus root (dr) 57.4% benjamin gilbert (f) 42.8%

Q: the majority of the district be new creation and do not have an incumbent

A: To determine if the majority of districts were newly created and did not have an incumbent, let's analyze the "incumbent" column:
1. **New York 1**: Incumbent: John Smith [...] 6. **New York 12**: Incumbent: David Thomas [...] From this analysis: - There is only one district (New York 1) that has an incumbent. [...]

(c) Incoherence



tie no	home team	score	away team	date
1	birmingham	2 - 0	port vale	24 january 1931
2	bury	1 - 2	exeter city	24 january 1931
3	watford	2 - 0	brighton & hove albion	24 january 1931
4	blackburn rovers	5 - 1	bristol rovers	24 january 1931
5	bolton wanderers	1 - 1	sunderland	24 january 1931
replay	sunderland	3 - 1	bolton wanderers	28 january 1931
6	grimsby town	1 - 0	manchester united	24 january 1931
7	west bromwich albion	1 - 0	tottenham hotspur	24 january 1931
8	sheffield united	4 - 1	notts county	24 january 1931
9	barnsley	2 - 1	sheffield wednesday	24 january 1931
10	brentford	0 - 1	portsmouth	24 january 1931
11	bradford city	0 - 0	wolverhampton wanderers	24 january 1931
replay	wolverhampton wanderers	4 - 2	bradford city	28 january 1931
12	crystal palace	0 - 6	everton	24 january 1931
13	chelsea	2 - 1	arsenal	24 january 1931
14	bradford park avenue	2 - 0	burnley	24 january 1931
15	southport	2 - 1	blackpool	24 january 1931
16	leeds united	4 - 1	newcastle united	24 january 1931

Q: of the 18 game play , 8 result in 1 or more team be shut out

A: To determine how many of the 18 games resulted in at least one team being shut out, we need to analyze each game's score:
1. Birmingham vs Port Vale: 2-0 (Birmingham shut out) 2. Bury vs Exeter City: 1-2 (Exeter City shut out) [...]

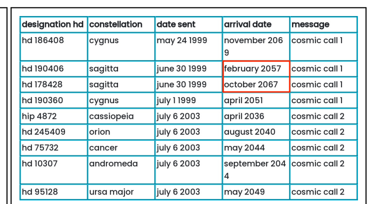
(d) Misunderstanding

week	date	opponent	result	attendance
1	september 7 1980	seattle seahawks	w 34 - 13	62042
2	september 14 1980	oakland raiders	w 30 - 24	51943
3	september 21 1980	denver broncos	w 30 - 13	74970
4	september 28 1980	kansas city chiefs	w 24 - 7	45161
5	october 5 1980	buffalo bills	l 26 - 24	51982
6	october 12 1980	oakland raiders	l 38 - 24	44826
7	october 19 1980	new york giants	w 44 - 7	50397
8	october 26 1980	dallas cowboys	l 42 - 31	60639
9	november 2 1980	cincinnati bengals	w 31 - 14	46406
10	november 9 1980	denver broncos	l 20 - 13	51435
11	november 16 1980	kansas city chiefs	w 20 - 7	50248
12	november 20 1980	miami dolphins	w 27 - 24	63013
13	november 30 1980	philadelphia eagles	w 22 - 21	51567
14	december 7 1980	washington redskins	l 40 - 17	48556
15	december 13 1980	seattle seahawks	w 21 - 14	49980
16	december 22 1980	pittsburgh steelers	w 26 - 17	51785

Q: the smallest crowd the charger play in front of be 44826, a road game against the raider

A: To determine if the smallest crowd the Chargers played in front of was 44,826 during a road game against the Raiders, let's follow these steps:
1. Identify the games where the Chargers played against the Raiders: [...] 2. Check the attendance for each game: [...]
3. Compare the attendance figures: [...]

(e) Faulty Methodology



designation hd	constellation	date sent	arrival date	message
hd 186408	cygnus	may 24 1999	november 2069	cosmic call 1
hd 190406	sagitta	june 30 1999	february 2057	cosmic call 1
hd 178428	sagitta	june 30 1999	october 2067	cosmic call 1
hd 190360	cygnus	july 1 1999	april 2051	cosmic call 1
hip 4872	cassiopeia	july 6 2003	april 2036	cosmic call 2
hd 245409	orion	july 6 2003	august 2040	cosmic call 2
hd 75732	cancer	july 6 2003	may 2044	cosmic call 2
hd 10307	andromeda	july 6 2003	september 2044	cosmic call 2
hd 95128	ursa major	july 6 2003	may 2049	cosmic call 2

Q: hd 178428 's arrival date be 10 year later than that of hd 190406

A: [...]
Since the difference is 10 years and 10 months, it is not exactly 10 years. Therefore, HD 178428's arrival date is not 10 years later than that of HD 190406.
Ground Truth: Yes

(f) Gray Area

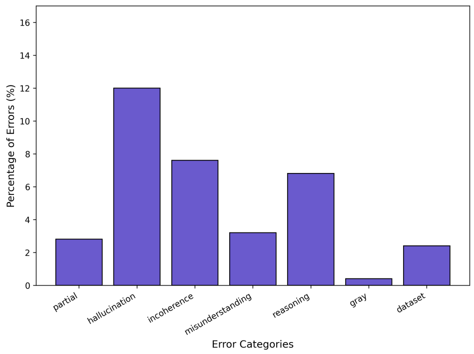
Figure 13: Illustration of the Error Categories

J Visual-TableQA Sample

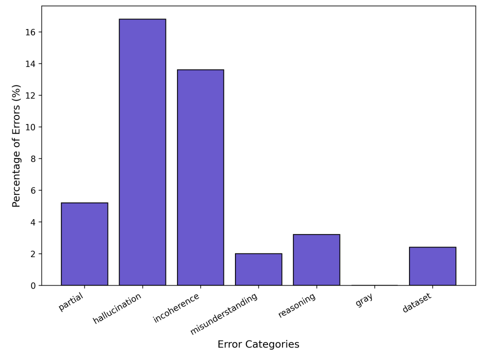
Table 10 gives some more detailed examples of our dataset samples.

K Image-to-LaTeX Dataset

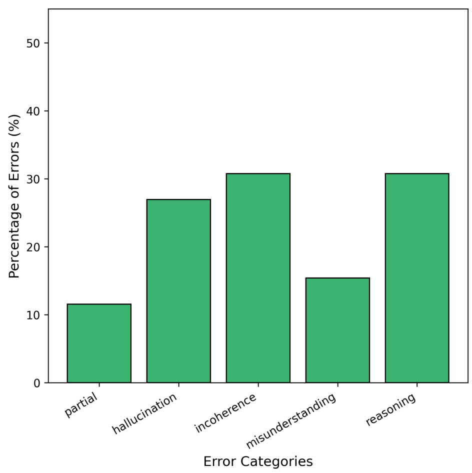
We have also released an additional dataset, **Img2TeX**, available on Hugging Face. It contains all the table images generated during the construction of **Visual-TableQA**, along with their corresponding



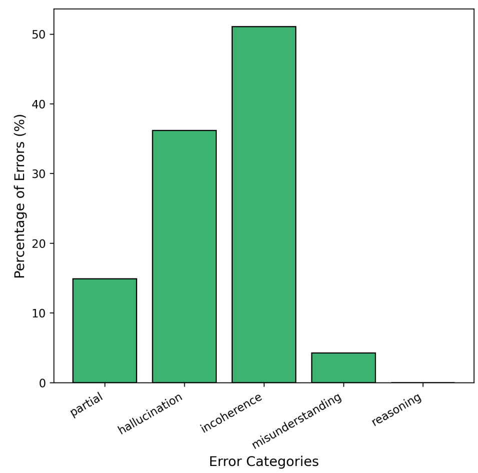
(a) Distribution of Errors for `Qwen2.5-vl-7b`.



(b) Distribution of Errors for `Qwen2.5-vl-7b` after Finetuning.



(c) Distribution of Errors corrected by Finetuning.

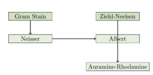


(d) Distribution of Errors introduced after Finetuning.

Figure 14: Comparison of error distributions between the pretrained `Qwen2.5-VL-7B` model (left) and its finetuned version (right). Percentages in the first row refer to the full *VTabFact* set. Percentages in the second row correspond to: (c) samples where the finetuned model was correct but the pretrained model was not, and (d) samples where the pretrained model was correct but the finetuned model was not.

LaTeX source code. This dataset is intended to complement the work of Kale & Nadadur (2025), which focuses on evaluating models' ability to generate LaTeX from textual prompts. In contrast, **Img2TeX** targets the task of generating LaTeX documents from visual (image-based) inputs. This distinction opens up a new evaluation pathway for vision-language models (VLMs), particularly those aiming to learn structured document generation from visual cues. As such, **Img2TeX** serves as a valuable benchmark for assessing the visual-to-structured-text generation capabilities of multimodal models.

Table 10: Sample of reasoning-intensive QA pairs. The first row's question and answer are truncated for readability. These questions address multiple visual aspects and extend beyond simple information extraction to test interpretive reasoning, as illustrated in the second row with a "how" question.

<table>
<tr><th>Table</th><th>Question</th><th>Answer</th></tr>
<tr><td>
Genealogy of Microbial Staining Methods by Risk Category and Application
<table>
<tr><th>Staining Method</th><th>Year Introduced</th><th>Primary Application</th><th>Risk Category</th><th>Key Reagent</th><th>Derivative Methods</th></tr>
<tr><td>Gram Stain</td><td>1884</td><td>Bacterial Classification</td><td>Low</td><td>Crystal Violet</td><td>Neisser, Hucker</td></tr>
<tr><td>Ziehl-Neelsen</td><td>1882</td><td>Acid-fast Bacteria</td><td>Moderate</td><td>Carbol Fuchsin</td><td>Kinyoun, Auramine-Rhodamine</td></tr>
<tr><td>Endospore Stain</td><td>1922</td><td>Spore Detection</td><td>Moderate</td><td>Malachite Green</td><td>Schaeffer-Fulton</td></tr>
<tr><td>Capsule Stain</td><td>1902</td><td>Capsule Visualization</td><td>Low</td><td>India Ink</td><td>Anthony, Maneval</td></tr>
<tr><td>Flagella Stain</td><td>1917</td><td>Motility Study</td><td>High</td><td>Tannic Acid</td><td>Leifson</td></tr>
<tr><td>Silver Stain</td><td>1937</td><td>Fungi / Protozoa</td><td>Very High</td><td>Silver Nitrate</td><td>Gomori, Warthin-Starry</td></tr>
<tr><td>Giemsa Stain</td><td>1904</td><td>Blood Parasites</td><td>Moderate</td><td>Eosin Azure</td><td>Wright, May-Grünwald</td></tr>
<tr><td>Acid-Fast Fluorochrome</td><td>1950</td><td>Mycobacteria</td><td>High</td><td>Auramine O</td><td>Auramine-Rhodamine</td></tr>
<tr><td>Periodic Acid-Schiff</td><td>1946</td><td>Fungal Cell Walls</td><td>High</td><td>Schiff Reagent</td><td>PAS-Diastase</td></tr>
<tr><td>Wright Stain</td><td>1902</td><td>Blood Smears</td><td>Moderate</td><td>Methylene Blue</td><td>Giemsa, May-Grünwald</td></tr>
<tr><td>Neisser Stain</td><td>1890</td><td>Corynebacteria</td><td>Low</td><td>Methylene Blue</td><td>Gram, Albert</td></tr>
<tr><td>Albert Stain</td><td>1921</td><td>Metachromatic Granules</td><td>Moderate</td><td>Toluidine Blue</td><td>Neisser</td></tr>
<tr><td>Auramine-Rhodamine</td><td>1950</td><td>Acid-fast Bacteria</td><td>High</td><td>Auramine O</td><td>Ziehl-Neelsen</td></tr>
<tr><td>Gomori Methenamine Silver</td><td>1953</td><td>Fungi Detection</td><td>Very High</td><td>Silver Nitrate</td><td>Silver Stain</td></tr>
<tr><td>PAS-Diastase</td><td>1957</td><td>Glycogen Removal</td><td>High</td><td>Diastase</td><td>PAS</td></tr>
</table>


Staining Method Notes
Note: Risk categories are based on reagent toxicity and laboratory safety. Derivative methods are direct adaptations or improvements of the original technique.
</td>
<td>Based on the visual patterns and row coloring in the table, what relationship exists between the alternating yellow-highlighted rows and the brewing methods' characteristics [...]?</td>
<td>Looking at the table's visual structure, the yellow-highlighted rows (rows 2, 4, 6, and 8) correspond to French Press, Chemex, Espresso, and Cold Brew methods respectively. These methods represent distinct extraction categories: French Press uses immersion, Chemex uses slow filtration, Espresso uses high-pressure extraction, and Cold Brew uses cold water extraction. The alternating color pattern appears to highlight methods that represent fundamental brewing principles or landmark techniques within each difficulty tier.[...]</td></tr>
<tr><td>
Table 1: Pioneering Film Transitions by Era
<table>
<tr><th>Transition Type</th><th>First Usage Year</th><th>Notable Films</th><th>Technical Diagram</th></tr>
<tr><td rowspan="3">Match Cut</td><td>1921</td><td>The Kid (Keaton)</td><td rowspan="3">Matched Object</td></tr>
<tr><td>1968 (modern use)</td><td>2001: A Space Odyssey</td></tr>
<tr><td>2010</td><td>Inception</td></tr>
<tr><td rowspan="2">Wipe</td><td>1916</td><td>The Cabinet of Dr. Caligari</td><td rowspan="2"></td></tr>
<tr><td>1977</td><td>Star Wars (light saber wipe)</td></tr>
<tr><td>Jump Cut</td><td>1920s (avant-garde)</td><td>Un Chien Andalou</td><td></td></tr>
<tr><td rowspan="4">Dissolve</td><td>1903</td><td>The Great Train Robbery</td><td rowspan="4"></td></tr>
<tr><td>1930s (standardized)</td><td>Gone with the Wind</td></tr>
<tr><td>1960s (soft focus)</td><td>Blowup</td></tr>
<tr><td>2000s (digital)</td><td>Inglourious Basterds</td></tr>
<tr><td rowspan="2">Crossfade</td><td>1910s</td><td>The Birth of a Nation</td><td rowspan="2"></td></tr>
<tr><td>2010s (dynamic)</td><td>Mad Max: Fury Road</td></tr>
</table>
</td>
<td>How does the genealogy of origami bases demonstrate a progressive complexity of folding techniques, and what key transformational steps are evident in the progression from a square sheet to more complex bases?</td>
<td>The genealogy reveals a systematic progression from simple to complex bases, starting with the fundamental Square Sheet. The Preliminary Fold introduces diagonal and perpendicular creases, serving as a critical intermediate step. This base enables more complex transformations like the Kite and Fish Bases, which develop pointed structures. The progression continues to the Bird Base, which combines multiple folding techniques to create more intricate models with distinct appendages. Each base builds upon previous folding techniques, demonstrating an incremental approach to increasing geometric complexity and model potential.</td></tr>
<tr><td>
Table 1: Interaction Matrix of Lighting Setups for Various Photographic Materials
<table>
<tr><th>Material</th><th>Diffused Light</th><th>Spotlight</th><th>Backlight</th><th>Side Light</th><th>Reflected Light</th></tr>
<tr><td>Silk</td><td>Soft shadows, smooth texture</td><td>Glare, loss of detail</td><td>Contour highlighting</td><td>Texture striations</td><td>Warm undertones</td></tr>
<tr><td>Glass</td><td>No distortion, clear refraction</td><td>Overly bright highlights</td><td>Silhouette definition</td><td>Prismatic effects</td><td>Moiré patterns</td></tr>
<tr><td>Metal</td><td>Even reflection control</td><td>Hotspots, glare</td><td>Edge luminosity</td><td>Surface texture visibility</td><td>Color shifting</td></tr>
<tr><td>Leather</td><td>Natural grain preservation</td><td>Overly contrasted creases</td><td>Depth enhancement</td><td>Grain texture emphasis</td><td>Shade discoloration</td></tr>
<tr><td>Wood</td><td>Grain visibility optimization</td><td>Ring-like highlights</td><td>Edge definition</td><td>Knot prominence</td><td>Color saturation loss</td></tr>
</table>
Color coding: Green = optimal, Red = problematic, Blue/Orange/Yellow = mixed results
</td>
<td>Which UNESCO site has the highest cultural impact indicator, and what symbolic meaning does it represent? How does this relate to its visitor numbers and year of inscription?</td>
<td>The Great Wall of China has the highest cultural impact indicator with a value of 120. It symbolizes 'Strength and perseverance' (as indicated by the blue color coding and symbolism type 2). Despite having the second-highest visitor numbers (10 million/year), it achieved maximum cultural impact. This suggests that while visitor numbers contribute to cultural impact, other factors like historical significance and symbolic representation play a stronger role. The site was inscribed in 1987, earlier than many others, giving it more time to establish its cultural influence.</td></tr>
</table>