Table Understanding and (Multimodal) LLMs: A Cross-Domain Case Study on Scientific vs. Non-Scientific Data

Ekaterina Borisova[1,2], **Fabio Barth**[1], **Nils Feldhus**[1,2,3],
Raia Abu Ahmad[1,2], **Malte Ostendorff**[4], **Pedro Ortiz Suarez**[5],
Georg Rehm[1,6], **Sebastian Möller**[1,2]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI),
[2]Technische Universität Berlin, [3]BIFOLD, [4]Deutsche Telekom,
[5]Common Crawl Foundation, [6]Humboldt-Universität zu Berlin
Corresponding author: ekaterina.borisova@dfki.de

Abstract

Tables are among the most widely used tools for representing structured data in research, business, medicine, and education. Although LLMs demonstrate strong performance in downstream tasks, their efficiency in processing tabular data remains underexplored. In this paper, we investigate the effectiveness of both text-based and multimodal LLMs on table understanding tasks through a cross-domain and cross-modality evaluation. Specifically, we compare their performance on tables from scientific vs. non-scientific contexts and examine their robustness on tables represented as images vs. text. Additionally, we conduct an interpretability analysis to measure context usage and input relevance. We also introduce the **TableEval** benchmark, comprising 3017 tables from scholarly publications, Wikipedia, and financial reports, where each table is provided in five different formats: Image, Dictionary, HTML, XML, and LaTeX. Our findings indicate that while LLMs maintain robustness across table modalities, they face significant challenges when processing scientific tables.

1 Introduction

Tables are one of the most ubiquitous tools for presenting data in a structured or semi-structured manner. They are commonly represented in a variety of textual (e. g., HTML, LaTeX, XML) or image formats (e. g., PNG, JPEG) and used across domains such as finance, medicine, and business, as well as in research and education.

In recent years, there has been a growing interest in table understanding (TU) techniques (Zhang and Balog, 2020; Gorishniy et al., 2021; Sahakyan et al., 2021; Borisov et al., 2022; Sui et al., 2024; Deng et al., 2024), aiming to extract and interpret information and knowledge contained in tables for tasks such as question answering (QA) and table-to-text generation (T2T) (Nan et al., 2022; Cheng et al., 2022; Osés Grijalba et al., 2024; Zheng et al., 2024). While large language models (LLMs) demonstrate strong performance in a wide range of applications (Chang et al., 2024; Raiaan et al., 2024; Caffagni et al., 2024; Zhang et al., 2024a; Team et al., 2024; OpenAI et al., 2024), their ability to understand (semi-)structured data remains under-researched (Sui et al., 2024; Fang et al., 2024) – especially for tables from *scientific* sources such as peer-reviewed articles, conference proceedings, and pre-prints.[1] There is also limited research on the impact of the representation modality of structured data (i. e., image vs. text) on model performance (Deng et al., 2024; Zhang et al., 2024d), and to the best of our knowledge, there are no approaches yet that specifically address scientific tables. In particular, most TU studies primarily focus on tables from *non-scientific* contexts such as Wikipedia (Parikh et al., 2020; Chen et al., 2021; Marzocchi et al., 2022; Wu et al., 2024b; Pang et al., 2024). However, compared to these domains, scientific tables often include technical terminology, complex concepts, abbreviations, and dense numerical values, requiring domain-specific knowledge and strong arithmetic reasoning skills (Ho et al., 2024; Moosavi et al., 2021). Recent works (Yang et al., 2025; Wu et al., 2024a) indicate that scientific tables present challenges to multimodal LLMs (MLLMs) and incorporating such (semi-)structured data into pretraining improves performance. As the number of published articles continues to increase rapidly (Fortunato et al., 2018; Bornmann et al., 2021; Hong et al., 2021), TU for scientific contexts, e. g., for scholarly document processing including information extraction and research knowledge graph construction, is becoming even more relevant. Finally, we

[1]Throughout this paper, we refer to such tables as *scientific* and to tables from other sources as *non-scientific*.

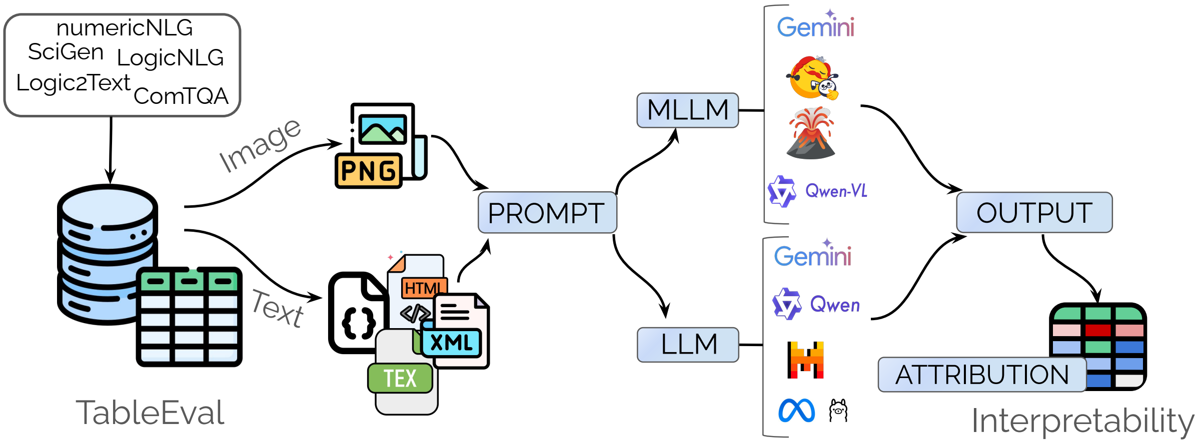

Figure 1: Schematic representation of the main phases in our experiments: 1. Develop TableEval dataset, 2. Evaluate each (M)LLM on individual data subsets from TableEval using various table representations (Image, LaTeX, XML, HTML, Dict), 3. Apply interpretability tools to the output yielding post-hoc feature attributions (e. g., using gradient-based saliency) which signify the importance of each token with respect to the model's output.

notice that interpretability analysis (Ferrando et al., 2024) for TU has received little attention and remains underexplored (Fang et al., 2024).

In this paper, we address the aforementioned gaps by examining the efficiency of both LLMs and MLLMs on a set of TU tasks. Specifically, we compare their ability to handle (semi-)structured data from scientific and non-scientific sources and explore the effects of image vs. diverse text-based table representations on model performance. We also conduct feature importance analyses to interpret the use of context information in LLMs. Figure 1 illustrates the main phases of our experiments.

Our contributions can be summarised as follows:

- We introduce TableEval, a cross-domain benchmark containing 3017 tables from scholarly publications, Wikipedia, and financial reports, available in image and four text formats (Dictionary, HTML, XML, and LaTeX). The dataset is publicly available on Hugging Face: `https://huggingface.co/datasets/katebor/TableEval`

- We conduct an extensive evaluation revealing that, although current (M)LLMs remain robust across table modalities, their performance significantly declines on scientific tables compared to non-scientific ones.

- We examine the applicability of gradient-based explanations for LLMs (Sarti et al., 2023) to TU to learn about the relevance of table content in prompts.

2 TableEval benchmark

Since no existing dataset covers both scientific and non-scientific tables across text and image modalities, we construct a benchmark tailored to our evaluation. This section outlines the collection processes of data (§2.1) and diverse table formats (§2.2).

2.1 Source data

To study the cross-domain performance of (M)LLMs, we developed the TableEval benchmark by leveraging pre-existing datasets of scientific and non-scientific tables. We collected relevant datasets based on the following criteria: 1. data is open-access; 2. test set with the gold labels is available; 3. metadata includes references to the sources of tables, such as DOIs for scholarly papers or URLs for Wikipedia pages; 4. target tasks (e. g., QA, T2T) are identical or very similar across datasets to maintain consistency and ensure comparability; 5. tables can be converted to the pre-defined formats (see §2.2). The following five datasets were selected (see Table 1): (a) **ComTQA** (Zhao et al., 2024), a visual QA (VQA) benchmark containing tables from PubTables-1M (Smock et al., 2022) and FinTabNet (Zheng et al., 2020), originating from PubMed Central[2] (PMC) papers and annual earnings reports, respectively. The annotations are generated using Gemini Pro (Team et al., 2024) and include questions requiring multiple answers, calculations, and logical reasoning. (b) **numericNLG** (Suadaa et al., 2021), a dataset focusing on the T2T generation task with numerical reasoning based on tables and

[2] `https://pubmed.ncbi.nlm.nih.gov`

Dataset	Task	Source	Image	Dict	LaTeX	HTML	XML
Scientific tables							
ComTQA (PubTables-1M)	VQA	PubMed Central	⬇	⚙	⚙	⚙	⎘
numericNLG	T2T	ACL Anthology	⎘	⬇	⚙	⬇	⚙
SciGen	T2T	arXiv and ACL Anthology	⎘	⬇	⎘	⚙	⚙
Non-scientific tables							
ComTQA (FinTabNet)	VQA	Earnings reports of S&P 500 companies	⎘	⚙	⚙	⚙	⚙
LogicNLG	T2T	Wikipedia	⚙	⬇	⚙	⎘	⚙
Logic2Text	T2T	Wikipedia	⚙	⬇	⚙	⎘	⚙

Table 1: Overview on the formats and collection methods for each dataset. Symbol ⬇ indicates formats already available in the given corpus, while ⎘ and ⚙ denote formats extracted from the table source files (e. g., article PDF, Wikipedia page) and generated from other formats in this study, respectively.

their textual descriptions extracted from ACL Anthology[3] articles and annotated by experts in the Computer Science field. (c) **SciGen** (Moosavi et al., 2021), a corpus designed for reasoning-aware T2T generation, comprising tables from arXiv[4] papers across fields such as Computation and Language, Machine Learning, Computer Science, Computational Geometry, etc. Its test set contains expert-annotated data. (d) **LogicNLG** (Chen et al., 2020a), a T2T dataset of open-domain tables from Wikipedia and associated with manually annotated natural language statements that can be logically entailed by the given data. (e) **Logic2Text** (Chen et al., 2020c), features open-domain Wikipedia tables manually annotated with descriptions of common logic types and their underlying logical forms for the T2T task. As shown in Table 1, the final TableEval corpus contains six data subsets, covering two downstream tasks (QA and T2T), and comprising 3017 tables and 11312 instances in total (for the detailed statistics see Table 4 in Appendix A). All annotations are taken from the source datasets. Examples from each dataset are provided in Appendix B.

2.2 Table formats

We represent tables from each TableEval subset as PNG images and in structured or semi-structured textual formats including HTML, XML, LaTeX, and Python Dictionary (Dict) to analyse LLMs' performance across different modalities. HTML is chosen as it is the original format of Wikipedia tables, XML for its use in encoding tables from PMC articles, LaTeX as it is the primary format for scientific tables, and Dict since it is readily available in most source datasets. Instances of tables in various representation formats were obtained using one of the following methods (see Table 1): 1. extraction from the original dataset; 2. extraction from the table source (e. g., article PDF); 3. generation from other formats (e. g., HTML ⇔ XML). Note that for the latter two, we manually validate the final results for each format and data subset by checking a random sample of about 100 instances. In what follows, the way we assembled each table format in the TableEval corpus is described in detail. Additional information is provided in Appendix C.

Image. Since the PubTables-1M subset of ComTQA already includes JPGs of tables, we simply convert them to PNGs. In contrast, other datasets provide only textual representations of tables. Thus, for numericNLG and SciGen, we first collect PDF files of the arXiv and ACL papers, and then use the PDFFigure2.0 (Clark and Divvala, 2016) tool to extract images of tables.[5] Whenever PDFFigure2.0 fails to produce an image, we utilise the MinerU tool (Wang et al., 2024) as an alternative. Note that SciGen instances associated with papers that are no longer open-access or do not contain tables are excluded. In case of FinTabNet, images of tables are extracted from the corresponding PDF pages of financial reports using the gold annotations of the bounding boxes. Finally, images of the Wikipedia tables in LogicNLG and Logic2Text are generated by converting their HTML representations into PNG files with the imgkit Python wrapper[6]. Distribution of image aspect rations across data subsets is provided in Figure 12 in Appendix D.

XML and HTML. PubTables-1M is the only dataset where the original XML sources of tables

[3] https://aclanthology.org
[4] https://arxiv.org

[5] In SciGen, some PDFs are taken from the ACL Anthology as they are no longer available on arXiv.
[6] https://pypi.org/project/imgkit/

can be obtained. To achieve this, we retrieve the source papers based on their PMC ID using the E-utilities API[7] and extract the tables with the ElementTree parser[8]. When it comes to HTML, we are unable to retrieve the original format since systematic downloading of article batches from the PMC website is prohibited[9]. This is why we generate HTML from XML using a custom Python script instead. Similarly, for numericNLG, we convert already available HTML into XML with a Python script. For SciGen, we download the source LaTeX code of each paper from arXiv, use the LaTeXML tool[10] to produce both XML and HTML, and extract tables from the resulting files. In contrast, we construct HTML for FinTabNet tables by leveraging gold annotations of HTML structure which provide tags and associated cell values. Afterwards, the HTML code is converted to XML in the same way as described for numericNLG. Finally, HTML in LogicNLG and Logic2Text are collected from the respective Wikipedia pages, while the XML format is obtained using the same approach applied to numericNLG and FinTabNet.

LaTeX. For SciGen, we obtain the LaTeX code directly from the source files of the papers. In contrast to arXiv data, no LaTeX code is available for PMC and ACL papers. Thus, we generate LaTeX for numericNLG and PubTables-1M tables from their HTML representations. To ensure the validity of the output, we compile the code and resolve any errors encountered. The same approach is used to obtain LaTeX for Wikipedia and financial tables.

Dictionary. All datasets except ComTQA already include linearised tables represented as lists of column headers and cell values, although the encoding conventions slightly vary across them (see Appendix C). To align with these datasets, we collect column headers, subheaders, and cell values for the PMC subset in ComTQA by parsing the table XML code with ElementTree. In case of FinTabNet, we extract these elements from a dataframe representation of each table obtained during the HTML collection phase. For the experiments, the linearised tables are represented as a Dict containing lists of column headers, lists of subheaders (if extracted), lists of rows, as well as title, caption, and footnote (if available).

3 Experiments

We benchmark various (M)LLMs using individual data subsets and representations of tables from TableEval. This is followed by an interpretability analysis applied to the output yielding attributions from a gradient-based method. In the following, we first describe the experimental set up (§3.1), then report and analyse the results (§3.2).

3.1 Experimental setup

Models. We evaluate both smaller and larger models in terms of parameter size (3-14 billion), see Table 2.[11] We primarily focus on open-source instruction-tuned (M)LLMs published on Hugging Face[12] (HF). The only closed-source model we use is Gemini-2.0-Flash (Team et al., 2024), which serves as our baseline, since Gemini is currently considered among the state-of-the-art. For MLLMs, we select LLaVa-NeXT (Li et al., 2024), Qwen2.5-VL (Bai et al., 2025), and Idefics3 (Laurençon et al., 2024). As for text-based LLMs, we evaluate Llama-3 (Grattafiori et al., 2024), Qwen2.5 (Qwen et al., 2025), and Mistral-Nemo[13].

Model	HF checkpoint	Size (B)	Vision
Gemini-2.0-Flash	–	–	✔
LLaVa-NeXT	llama3-llava-next-8b-hf	8	✔
Qwen2.5-VL	Qwen2.5-VL-3B-Instruct	3	✔
	Qwen2.5-VL-7B-Instruct	7	✔
Idefics3	Idefics3-8B-Llama3	8	✔
Llama-3	Llama-3.2-3B-Instruct	3	✖
Qwen2.5	Qwen2.5-3B-Instruct	3	✖
	Qwen2.5-14B-Instruct	14	✖
Mistral-Nemo	Mistral-Nemo-Instruct-2407	12	✖

Table 2: (M)LLMs used in the experiments ("Size" indicates the number of parameters in billions).

Prompts and data. We run experiments on every data subset from the TableEval corpus and develop prompt templates that are customised to each task, applying them uniformly across all models to ensure consistency during the evaluation. To study the models' true capability to understand various table representations, we exclude explicit document type indicators (e. g., HTML/XML headers) and do not specify the format in the prompt. Additionally, given the diversity of the (M)LLMs and the fact that they may not always adhere to a specific

[7] https://www.ncbi.nlm.nih.gov/home/develop/api/

[8] https://docs.python.org/3/library/xml.etree.elementtree.html#

[9] https://pmc.ncbi.nlm.nih.gov/about/copyright/

[10] https://math.nist.gov/~BMiller/LaTeXML/

[11] Due to limited computational resources, we restricted the evaluation to (M)LLMs with up to 14 billion parameters.

[12] https://huggingface.co

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

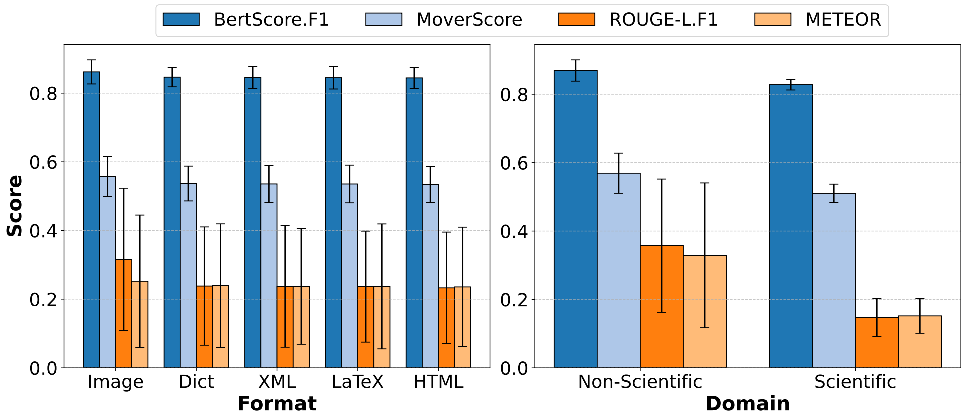

Figure 2: BertScore.F1, MoverScore, ROUGE-L.F1, and METEOR for the table formats averaged over data subsets and models (left), and for scientific vs. non-scientific domain averaged over data subsets, models, and formats (right). Error bars indicate standard deviation.

output structure (which can hinder proper parsing of the answer), we do not enforce a particular response format. The prompt templates are provided in Appendix E.

Evaluation metrics. We follow the scores reported in the original papers for each data subset. Thus, we compute BLEU-N (Papineni et al., 2002), SacreBLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005), ROUGE-N, ROUGE-L (Lin, 2004), MoverScore (Zhao et al., 2019), BertScore (Zhang* et al., 2020), and BLEURT (Sellam et al., 2020). Given the extensive set of metrics, we report only BertScore.F1, MoverScore, ROUGE-L.F1, and METEOR in the main text, while providing all raw score values in Appendix F.

Interpretability analysis. Inseq (Sarti et al., 2023) applies feature attribution methods to generative LLMs to highlight how important each token in the input is for generating the next token with the help of a heatmap. In our experimental setup, we perform post-hoc analyses using the model outputs as custom attribution targets on an instance level. Input x Gradient (Simonyan et al., 2014), provided by Inseq, is selected as it is both computationally efficient and more faithful than, e. g., attention weights. The saliency is averaged to produce a one-dimensional vector of token attributions, which we visualise as a heatmap.

Implementation details. All experiments are conducted in a zero-shot setting using the (M)LLMs' default hyperparameters with the seed value set to 42. We choose the batch size equal to 1 for all open-source (M)LLMs and to the size of the given subset for Gemini-2.0-Flash. We use Nvidia A100 (40GB, 80GB), H100 (80GB), H200 (141GB), and L40S (48GB) GPUs for the open-source models depending on the given LLM and TableEval subset size. The Gemini-2.0-Flash results are evaluated using the Batch API through the LiteLLM framework[14]. We developed an end-to-end evaluation pipeline[15] for the experiments and use HF transformers or LiteLLM and the datasets library to load the models and datasets, respectively.

3.2 Results and analysis

Image vs. text. Averaged score values across models and data subsets for each table format are given in Figure 2 (left), whereas raw results are shown in Table 5 in Appendix F. The use of images outperforms the use of text across all metrics by approximately 1-13%. In particular, for ComTQA and LogicNLG, image achieves the best results, while for other data subsets the outcomes are either similar or the text modality prevails (by about 1-10%), as shown in Figure 3 a) and Tables 6–11 in Appendix F. This aligns with previous studies (Deng et al., 2024) reporting comparable or significantly better performance of models on the vision modality. Unlike prior works (Sui et al., 2024; Singha et al., 2023; Deng et al., 2024), we do not observe a large variation in results across LLMs and the four text formats, with the maximum gap equal to about 4%. Further analysis of the metrics for individual models and formats also indicates similar accuracy across the LLMs, see Figure 3 b) and Tables 12–16 in Appendix F. Hence, our find-

[14] https://www.litellm.ai
[15] https://github.com/esborisova/TableEval-Study

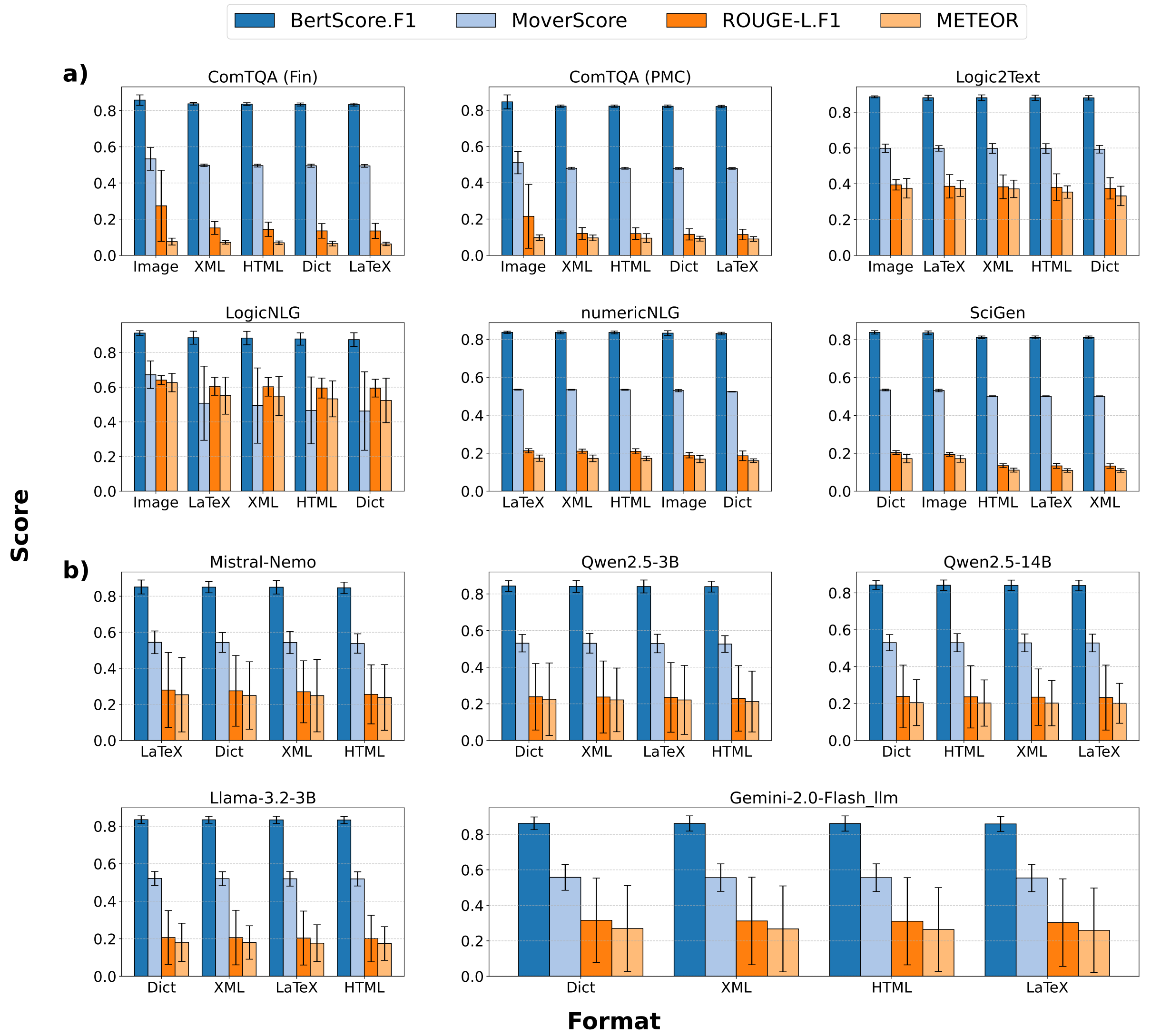

Figure 3: Values of BertScore.F1, MoverScore, ROUGE-L.F1, and METEOR **a)** for individual data subsets and all formats averaged over models, and **b)** for individual models and text formats averaged over data subsets. Error bars indicate standard deviation. Here "Fin" stands for FinTabNet, "PMC" denotes PubTables-1M, while "_llm" indicates text input for Gemini-2.0-Flash.

ings suggest that current models are less sensitive to diverse text representations of tables. Such outcomes may be attributed to LLMs' exposure to data encoded in the given formats during pretraining.

Scientific vs. non-scientific. The results for each domain are shown in Figure 2 (right) and Table 17 in Appendix F. The findings indicate that LLMs are more efficient on TU tasks from the non-scientific split, achieving a score boost of up to 34%. The best score values are obtained for LogicNLG followed by Logic2Text, see Figure 4 (left) and Table 18 in Appendix F.

We hypothesise that this difference could arise from (a) the complexity level of the given data and the target task; (b) lack or sparsity of the data from scientific contexts in the pre-training corpus of (M)LLMs. In numericNLG and SciGen, the goal is to generate a coherent paragraph or a collection of paragraphs summarising the table's content. In contrast, both LogicNLG and Logic2Text involve producing a single statement, filling in masked entities in a sentence and generating text based on a logical form, respectively. Furthermore, according to Moosavi et al. (2021), SciGen is characterised by a higher level of complexity than LogicNLG. This is because each gold description in SciGen summarises the entire table content and involves multiple types of reasoning, whereas, in LogicNLG each statement often focuses on a subset of table rows and is associated with a single type of reasoning. Similar to LogicNLG, Logic2Text

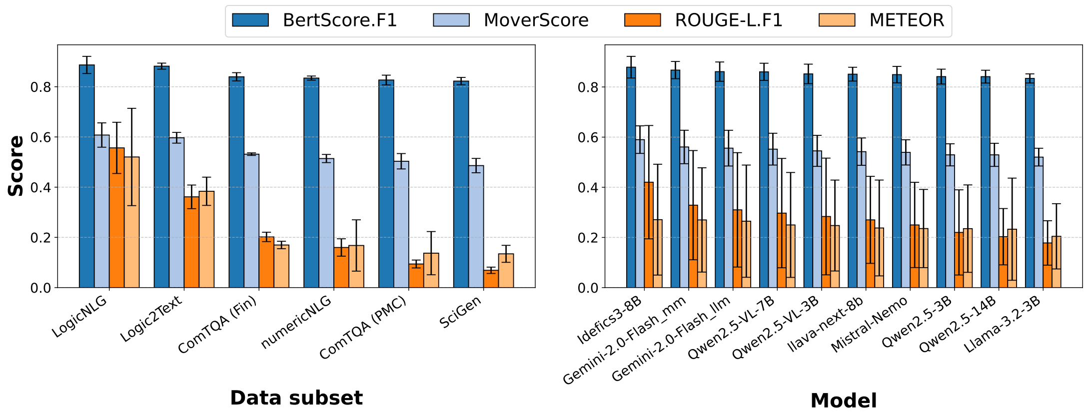

Figure 4: BertScore.F1, MoverScore, ROUGE-L.F1, and METEOR for each data subset averaged over table formats and models (left), and for individual models averaged over data subsets and formats (right). Error bars indicate standard deviation. Here "Fin" stands for FinTabNet, "PMC" denotes PubTables-1M, while "_llm" and "_mm" are used to distinguish between text and image input for Gemini-2.0-Flash, respectively.

descriptions involve only one type of logic. Notably, comparable performance is achieved across models for both subsets in ComTQA, with the gap in scores equal to about 1-3% (except for a 17% higher BLEURT score for PubTables-1M). Given that ComTQA was also proposed as a more challenging benchmark compared to existing datasets, comprising questions with multiple answers, numerical, and logical reasoning, the lower performance of (M)LLMs could lie in the complexity of the data as well. Finally, reasoning over scientific tables requires in-domain knowledge, the absence of which likely contributes to a decline in accuracy for the respective TableEval subsets.

Comparison of (M)LLMs. Figure 4 (right) and Table 19 in Appendix F outline results for individual models. Among MLLMs, Gemini-2.0-Flash and Idefics3 perform best, with the former outperforming the latter on BLEU-N, BLEURT, METEOR, ROUGE-3, and ROUGE-4 (by 1-4%). Next in the ranking are Qwen2.5-VL models and LLaVa-NeXT. For LLMs, Gemini-2.0-Flash obtains the highest score values, followed by Mistral-Nemo. Qwen2.5 models rank next with the 3B version achieving either similar or slightly better results than its 14B counterpart. On the contrary, Llama-3 consistently shows the weakest performance. We observe that on average, Idefics3 tends to generate concise responses with the shortest outputs produced for QA task (e. g., just a numeric value), whereas other models provide longer outputs. A similar trend is observed for LLMs, with Gemini-2.0-Flash providing shorter predictions compared to other models. Table 3 outlines the statistics on prediction lengths for each (M)LLM. Additionally, Figure 15 (Appendix F) illustrates the mean lengths for each model and data subset, while Figure 16 (Appendix G) demonstrates prediction examples. Since we do not postprocess the models' outputs, such difference in response length can contribute to the discrepancy across (M)LLMs in BLEU-N and ROUGE-N, which rely on n-gram overlap. Overall, our evaluation indicates that open-source models still remain behind the closed-source Gemini-2.0-Flash. On another note, we could not observe any correlation between model size and accuracy.

Model	**Mean**	**Min**	**Max**
Idefics3-8B-Llama3	139	0	4416
Qwen2.5-VL-3B-Instruct	360	2	4170
Qwen2.5-VL-7B-Instruct	292	4	3464
llama3-llava-next-8b-hf	311	24	6336
Gemini-2.0-Flash_mm	207	2	3097
Gemini-2.0-Flash_llm	259	0	10282
Llama-3.2-3B-Instruct	464	22	5626
Mistral-Nemo-Instruct-2407	303	21	2941
Qwen2.5-14B-Instruct	481	29	4154
Qwen2.5-3B-Instruct	465	26	4535

Table 3: Statistics on the mean, minimum, and maximum prediction lengths (in characters) for each model across TableEval subsets. Blue and pink colours highlight the lowest and highest values in each column, respectively. Here "_llm" and "_mm" are used to distinguish between text and image input for Gemini-2.0-Flash, respectively.

Interpretability. We choose instance-level analysis because dataset-level statistics tend to flatten important nuances, especially in generative settings

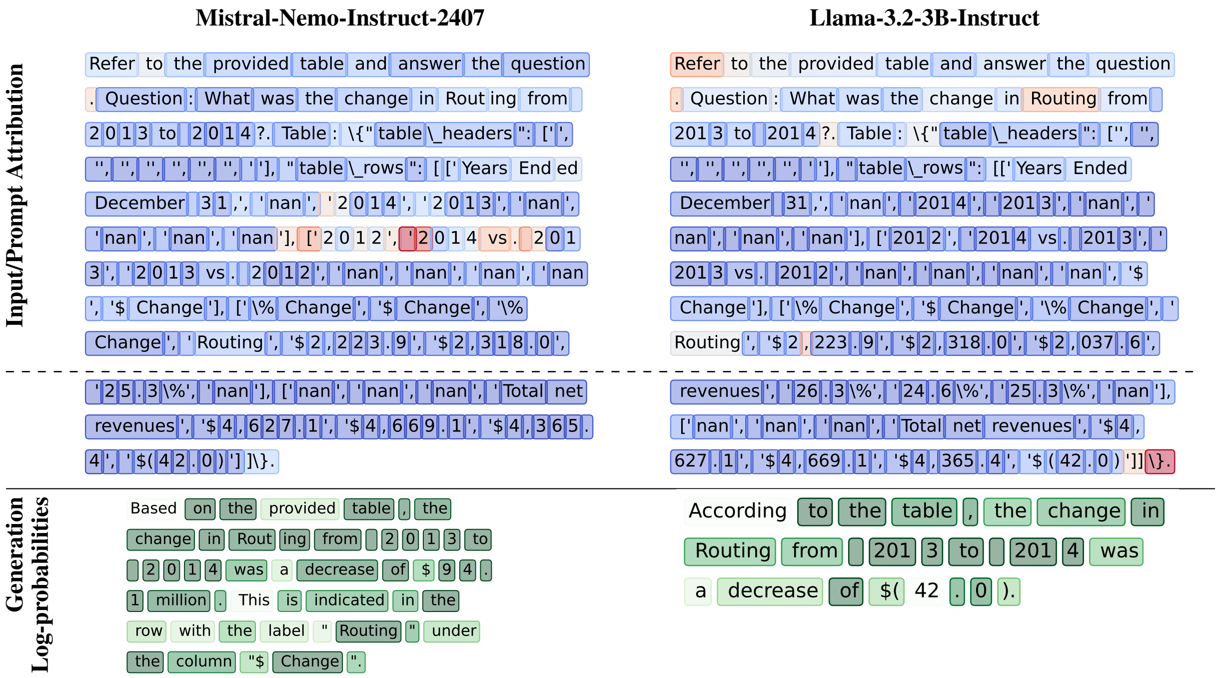

Figure 5: Interpretability analysis using Input x Gradient on Mistral-Nemo (correct prediction) and Llama3 (incorrect prediction) for a ComTQA (FinTabNet) instance with the Dict format. The gold answer to the given question is *"decrease of $94.1"*. Redder highlights correspond to higher importance. The prompts are abbreviated in the middle, indicated with the dashed line. In addition, for the output, we visualise the log-probabilities representing the model's confidence (dark green = very confident).

without a finite number of classes (Rönnqvist et al., 2022). Due to computational and visualisation constraints, we selected four ComTQA and two LogicNLG instances. The former was chosen for its shorter reference and prediction lengths compared to other subsets, while the latter was selected for achieving the highest scores across LLMs. We compare the best (Mistral-Nemo) and worst (Llama3) performing open-source LLMs.[16]

Figure 5 shows saliency maps as determined by the Input x Gradient explainer and log-probabilities for the generation (see §3.1). In this ComTQA (FinTabNet) example, with the table represented as a Dict in the input, we first notice that positive attributions are generally sparse due to the saturation problem (Shrikumar et al., 2017) and potentially the long context. Llama3 puts most attribution towards start and end of the prompt and the row value mentioned in the question (*"Routing"*). Mistral-Nemo, on the other hand, focuses much more on the year columns that are relevant to answering the question correctly. A key difference also lies in the tokenisation: While Mistral-Nemo splits all numbers into single digits, Llama3 often uses three-digit tokens where the fourth digit of a year is cut off. We assume that this makes it harder for Llama3 to process the marginal differences correctly.

The log-probabilities for the generated tokens are a proxy for the model's confidence. Here, we observe high uncertainty in Llama3 generating the core of the answer, the number token *"42"*, which is incorrect. Mistral-Nemo, on the contrary, correctly answers the question and we can see that it is certain about it from the high log-probabilities. Additionally, the model shows high confidence in the row *"Routing"* and column *"Change"* as the location of the answer, which indeed corresponds to the true position of the value (see also Figure 22 in Appendix H). At the same time, it is uncertain about optional, meaning-preserving generations such as the token *"provided"* as a qualifier for *"table"* and the beginning of the second sentence following the answer which serves as a rationale for the model's decision-making (Lu et al., 2024).

Appendix H shows five more examples for ComTQA and LogicNLG instances. We also observe a repeating pattern of the start and end of a prompt being attributed the most. While these observations are based on a small set of instances, our pipeline enables computing saliency maps for

[16] Saliency maps for these examples, along with additional instances, are available also in our GitHub repository.

any combination of prompt, input format, model, and dataset in future experiments.

4 Related work

Earlier TU studies leverage LLMs by representing tables as sequential text, either through naïve linearisation or by incorporating delimiters and special tokens (Fang et al., 2024). Some works focus on fine-tuning LLMs to enhance TU (Zhang et al., 2024c,b; Herzig et al., 2020; Yin et al., 2020; Gong et al., 2020; Iida et al., 2021), while others explore LLMs' table reasoning abilities through prompt engineering (Zhao et al., 2023; Chen, 2023; Sui et al., 2024). However, compared to natural language, tables present unique challenges to LLMs due to their varying layout structures, feature heterogeneity, and a large number of components leading to excessively long sequences (Borisov et al., 2022). The latter is particularly problematic, as most LLMs become inefficient due to the quadratic complexity of self-attention (Vaswani et al., 2017). With recent advances in vision and multimodality research, using MLLMs for TU has gained increasing attention with models like GPT-4 (OpenAI et al., 2024) and Gemini (Team et al., 2024), being widely adopted. Although, similar to LLMs, MLLMs also struggle with understanding structured data (Zheng et al., 2024).

Several studies examine the impact of the table representation on models' efficiency, indicating that different table formats suit specific TU tasks and LLMs at hand (Deng et al., 2024; Sui et al., 2024; Zhang et al., 2024d; Singha et al., 2023). For instance, Sui et al. (2024) find HTML and XML being better understood by GPT models than Markdown, JSON, and natural language with separators encoding. In contrast, Singha et al. (2023) observe that using HTML leads to lower performance for the fact-finding and transformation tasks compared to dataframe-based and JSON formats. Meanwhile, Deng et al. (2024) analyse how models' reasoning abilities vary when tables are represented as text vs. images showing that Gemini Pro and GPT-4 perform similarly across both modalities.

While these studies offer insights into the effectiveness of (M)LLMs in interpreting structured data across formats, they focus primarily on non-scientific contexts like Wikipedia and finance. This is likely due to the abundance of established, large-scale datasets based on tables from these sources, including WikiTables (Bhagavatula et al., 2015), ToTTo (Parikh et al., 2020), and TabFact, (Chen et al., 2020b), to name a few. Furthermore, interpretability for TU tasks remains under-researched, as related works mainly consider unstructured text and are disconnected from downstream applications (Ferrando et al., 2024; Tenney et al., 2024), rarely focusing on other long-form tasks like retrieval-augmented generation (Qi et al., 2024) or QA (Enouen et al., 2024). Nguyen et al. (2025) use attributions to make tabular QA explainable but they are constrained to the text-to-SQL setup. Unlike prior studies, this paper focuses on cross-domain and cross-modality evaluation, comparing the performance and explanations of (M)LLMs on both scientific and non-scientific tables, covering image and diverse text representations of tables.

5 Conclusion

We conducted an evaluation study to explore the robustness of diverse (M)LLMs on scientific vs. non-scientific tables across image and four text formats. The findings reveal that current models obtain decent performance across both vision and text modalities but significantly struggle with scientific tabular data. Additionally, we explored the applicability of interpretability methods to TU tasks to get insights into the decision-making of LLMs. We found feature attributions to be a useful tool for revealing model uncertainty, its attention to table structure and relevant content, and tokenisation differences which might potentially affect predictions.

Limitations

Although this study provides insights into the strengths and limitations of (M)LLMs in understanding tables, it has several limitations. First, we use the same prompts across (M)LLMs and do not postprocess the predictions which may contribute to lower score values. Experimenting with model-specific prompts and structured outputs using tools such as Jsonformer[17] could lead to better results. Second, we rely on automatic metrics, the drawbacks of which have been well-documented previously (Schmidtova et al., 2024; Gehrmann et al., 2023). Third, we focus only on interpretability for the text input, while methods like CC-SHAP (Parcalabescu and Frank, 2025) remain the next step to measure the importance of each modality in MLLM decision-making. Fourth, annotating all subsets in TableEval for a common task and

[17] https://github.com/1rgs/jsonformer

evaluating (M)LLMs on the entire corpus could be beneficial and we leave it for future work. Finally, the dataset is limited to the English language and thus does not allow for the assessment of multilingual TU.

Ethics statement

The data used in this study is based on publicly available datasets. We adhere to their respective licenses and conditions of use in our experiments. Additional table formats are generated with Python scripts and open-access tools or collected from the original table sources which are under permissive licenses. All (M)LLMs, except Gemini-2.0-Flash, employed for the experiments are open-access. Those models might potentially possess biases, as outlined by their developers, which researchers should be aware of.

Acknowledgments

This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[18] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The consortium is funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). We would like to thank Melina Plakidis, Maximilian Dustin Nasert, and Shuiai Xu for their help in manually reviewing certain subsets of the data.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL technical report. *Preprint*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. 2015. TabEL: Entity linking in web tables. In *The Semantic Web - ISWC 2015*, pages 425–441, Cham. Springer International Publishing.

Vadim Borisov, Tobias Leemann, Kathrin Sessler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

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

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W. Cohen. 2021. Open question answering over tables and text. *Preprint*, arXiv:2010.10439.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

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

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020c. Logic2Text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2096–2111, Online. Association for Computational Linguistics.

[18] https://www.nfdi4datascience.de

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

Christopher Clark and Santosh Divvala. 2016. PDF-Figures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 143–152, New York, NY, USA. Association for Computing Machinery.

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

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. 2024. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13984–14011, Bangkok, Thailand. Association for Computational Linguistics.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large language models (LLMs) on tabular data: Prediction, generation, and understanding – a survey. *Preprint*, arXiv:2402.17944.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *arXiv*, abs/2405.00208.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379):eaao0185.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. 2020. TableGPT: Few-shot table-to-text generation with table structure reconstruction and content matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1978–1988, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943. Curran Associates, Inc.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et. al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Xanh Ho, Anh Khoa Duong Nguyen, An Tuan Dao, Junfeng Jiang, Yuki Chida, Kaito Sugimoto, Huy Quoc To, Florian Boudin, and Akiko Aizawa. 2024. A survey of pre-trained language models for processing scientific text. *Preprint*, arXiv:2401.17824.

Zhi Hong, Logan Ward, Kyle Chard, Ben Blaiszik, and Ian Foster. 2021. Challenges and advances in information extraction from scientific literature: a review. *JOM*, 73:1543–1851.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. TABBIE: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456, Online. Association for Computational Linguistics.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xinyuan Lu, Liangming Pan, Yubo Ma, Preslav Nakov, and Min-Yen Kan. 2024. TART: An open-source tool-augmented framework for explainable table-based reasoning. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Mattia Marzocchi, Marco Cremaschi, Riccardo Pozzi, Roberto Avogadro, and Matteo Palmonari. 2022. MammoTab: A giant and comprehensive dataset for semantic table interpretation. In *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab2022)*.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. SciGen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Giang Nguyen, Ivan Brugere, Shubham Sharma, Sanjay Kariyappa, Anh Totti Nguyen, and Freddy Lecue. 2025. Interpretable LLM-based table question answering. *arXiv*, abs/2412.12386.

OpenAI, Josh Achiam, and Steven Adler et. al. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Jorge Osés Grijalba, L. Alfonso Ureña-López, Eugenio Martínez Cámara, and Jose Camacho-Collados. 2024. Question answering over tabular data with DataBench: A large-scale empirical evaluation of LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13471–13488, Torino, Italia. ELRA and ICCL.

Chaoxu Pang, Yixuan Cao, Chunhao Yang, and Ping Luo. 2024. Uncovering limitations of large language models in information seeking from tables. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1388–1409, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2025. Do vision & language decoders use images and text equally? How self-consistent are their explanations? In *The Thirteenth International Conference on Learning Representations*.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, Miami, Florida, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Samuel Rönnqvist, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter, and Veronika Laippala. 2022. Explaining classes through stable word attributions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1063–1074, Dublin, Ireland. Association for Computational Linguistics.

Maria Sahakyan, Zeyar Aung, and Talal Rahwan. 2021. Explainable artificial intelligence for tabular data: A survey. *IEEE Access*, 9:135392–135422.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of* the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in LLMs. *Preprint*, arXiv:2310.10358.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4624–4632.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465, Online. Association for Computational Linguistics.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets LLM: Can large language models understand structured table data? A benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA. Association for Computing Machinery.

Gemini Team, Rohan Anil, and Sebastian Borgeaud et. al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Ian Tenney, Ryan Mullins, Bin Du, Shree Pandya, Minsuk Kahng, and Lucas Dixon. 2024. Interactive prompt debugging with sequence salience. *arXiv*, abs/2404.07498.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. MinerU: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024a. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574, Bangkok, Thailand. Association for Computational Linguistics.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. 2024b. TableBench: A comprehensive and complex benchmark for table question answering. *Preprint*, arXiv:2408.09174.

Bohao Yang, Yingji Zhang, Dong Liu, André Freitas, and Chenghua Lin. 2025. Does table source matter? Benchmarking and improving multimodal scientific table understanding and reasoning. *arXiv*, abs/2501.13042.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Shuo Zhang and Krisztian Balog. 2020. Web table extraction, retrieval, and augmentation: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(2).

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024b. TableLlama: Towards open large generalist models for tables. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6024–6044, Mexico City, Mexico. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, Jifan Yu, Shu Zhao, Juanzi Li, and Jie Tang. 2024c. TableLLM: Enabling tabular data manipulation by LLMs in real office usage scenarios. *Preprint*, arXiv:2403.19318.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Baoxin Wang, Dayong Wu, Qingfu Zhu, and Wanxiang Che. 2024d. FLEXTAF: Enhancing table reasoning with flexible tabular formats. *arXiv*, abs/2408.08841.

Bowen Zhao, Changkai Ji, Yuejie Zhang, Wen He, Yingwen Wang, Qing Wang, Rui Feng, and Xiaobo Zhang. 2023. Large language models are complex table parsers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14786–14802, Singapore. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Shu Wei, Binghong Wu, Lei Liao, Yongjie Ye, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2024. TabPedia: Towards comprehensive visual table understanding with concept synergy. *Preprint*, arXiv:2406.01326.

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

Xinyi Zheng, Doug Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2020. Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. *Preprint*, arXiv:2005.00589.

A Dataset statistics

Dataset	Image		Dict		LaTeX		HTML		XML	
	Instances	**Tables**	**Instances**	**Tables**	**Instances**	**Tables**	**Instances**	**Tables**	**Instances**	**Tables**
Scientific tables										
ComTQA (PubTables-1M)	6232	932	6232	932	6232	932	6232	932	6232	932
numericNLG	135	135	135	135	135	135	135	135	135	135
SciGen	1035	1035	1035	1035	928	928	985	985	961	961
Total	**7402**	**2102**	**7402**	**2102**	**7295**	**1995**	**7352**	**2052**	**7328**	**2028**
Non-scientific tables										
ComTQA (FinTabNet)	2838	659	2838	659	2838	659	2838	659	2838	659
LogicNLG	917	184	917	184	917	184	917	184	917	184
Logic2Text	155	72	155	72	155	72	155	72	155	72
Total	**3910**	**915**	**3910**	**915**	**3910**	**915**	**3910**	**915**	**3910**	**915**

Table 4: Data distribution in the TableEval corpus for each format and subset.

B Dataset examples

QA task: ComTQA (PubTables-1M)

Table 5: Brood size analysis of *kin-29* alleles

Genotype	% of wild-type brood size
N2	100 (270)
sma-6(wk7)	64 (172)
lon-1(wk50)	81 (219)
kin-29(wk61)	32 (86)
kin-29(oy38)	81 (218)
kin-29(oy39)	80 (217)

Number of eggs scored for each genotype is shown in parentheses.

Question: What is the title of the table?

Answer: Brood size analysis of kin-29 alleles

Figure 6: An example from ComTQA (PubTables-1M), illustrating a table, a corresponding question, and a gold answer.

QA task: ComTQA (FinTabNet)

	Moody's	S&P	Fitch (a)
PPL Electric (b)			
Senior Unsecured/Issuer Rating	Baa1	A-	BBB
First Mortgage Bonds	A3	A-	A-
Senior Secured Bonds	A3	A-	A-
Commercial Paper	P-2	A-2	F2
Preferred Stock	Baa3	BBB	BBB
Preference Stock	Baa3	BBB	BBB
Outlook	STABLE	STABLE	STABLE

Question: What is the rating of commercial paper?

Answer: P-2 A-2 F2

Figure 7: An example from ComTQA (FinTabNet), illustrating a table, a corresponding question, and a gold answer.

T2T task: numericNLG

Genre	Sentences	Length	Yield	Precision
News*	100	19.3	142	78.9
News	100	19.3	144	70.8
Wiki	100	21.4	178	61.8
Web	100	19.2	165	49.1
Total	300	20.0	487	60.2

Table 1: Corpus size (length in token) and system performance by genre. News* used gold trees and is not included in total.

Description: Results. From the whole corpus of 300 sentences, PropsDE extracted 487 tuples, yielding on average 1.6 per sentence with 2.9 arguments. 60% of them were labeled as correct. Table 1 shows that most extractions are made from Wikipedia articles, whereas the highest precision can be observed for newswire text. According to our expectations, web pages are most challenging, presumably due to noisier language. These differences between the genres can also be seen in the precision-yield curve (Figure 2).

Figure 8: An example from numericNLG, illustrating a table and its corresponding gold description.

T2T task: SciGen

Model		Test	*but*	*but* or *neg*
no-distill	no-project	85.98	78.69	80.13
no-distill	project	86.54	83.40	-
distill [7]	no-project	86.11	79.04	-
distill	project	86.62	83.32	-
ELMo	no-project	88.89	86.51	87.24
ELMo	project	88.96	87.20	-

Table 2: Average performance (across 100 seeds) of ELMo on the SST2 task. We show performance on *A-but-B* sentences ("*but*"), negations ("*neg*").

Description: Switching to ELMo word embeddings improves performance by 2.9 percentage points on an average, corresponding to about 53 test sentences. Of these, about 32 sentences (60% of the improvement) correspond to A-but-B and negation style sentences, [CONTINUE] As further evidence that ELMo helps on these specific constructions, the non-ELMo baseline model (no-project, no-distill) gets 255 sentences wrong in the test corpus on average, only 89 (34.8%) of which are A-but-B style or negations.

Figure 9: An example from SciGen, illustrating a table and its corresponding gold description.

T2T task: LogicNLG

Country	Date	Label	Format	Catalogue No.
Europe	17 October 2008[160]	Columbia	CD, Double LP	#88697392232
Australia	18 October 2008[39]	Sony Music	CD	#88697392382
United Kingdom	20 October 2008[161] [162]	Columbia	CD, Double LP	#88697392232
	1 December 2008[38]		CD (limited edition steel-box)	#88697417452
United States	20 October 2008	Columbia	CD	#88697338292
Japan	22 October 2008[163]	Sony Music	CD	SICP-2055
Germany	5 December 2008[164]	Columbia	CD (limited edition steel-box)	#886974174523
Global (iTunes)	19 November 2012[49]	Columbia	Digital download	#88697338292

Title: black ice (album)

Template: the album [ENT] was first released in [ENT]

Statement: the album Black Ice was first released in Europe.

Figure 10: An example from LogicNLG, illustrating a table, a statement with masked entities, and a corresponding gold statement.

T2T task: Logic2Text

Pick #	CFL Team	Player	Position	College
13	Hamilton Tiger-Cats	Devin Grant	OL	Utah
14	BC Lions (via Winnipeg)	Matt Kellett	K	Saskatchewan
15	Montreal Alouettes (via Winnipeg via BC)	Scott Flory	OL	Saskatchewan
16	Calgary Stampeders	Harland Ah You	DL	Brigham Young
17	Edmonton Eskimos	Scott Deibert	RB	Minot State
18	Montreal Alouettes	William Loftus	D	Manitoba
19	Saskatchewan Roughriders	Kevin Pressburger	LB	Waterloo
20	Toronto Argonauts	Jermaine Brown	RB	Winona State

Title: `1998 cfl draft`

Logical form: `and { only { filter_eq { filter_eq { all_rows ; college ; saskatchewan } ; position ; k } } ; eq { hop { filter_eq { filter_eq { all_rows ; college ; saskatchewan } ; position ; k } ; player } ; matt kellett } } = true`

Statement: `the only kicker drafted by saskatchewan college in the 1998 cfl draft was matt kellett .`

Figure 11: An example from Logic2Text, illustrating a table, a logical form, and a corresponding gold statement.

C Table formats collection

In what follows, we provide additional details on the collection process of the table formats.

XML and HTML. As was mentioned in §2.2, XML and XML/HTML for the PubTables-1M subset of ComTQA and SciGen, respectively, are extracted from the source papers. For the former, the target tables are identified based on their titles and the highest cosine similarity with table content annotations available in PubTables-1M. For Scigen we use the fuzzy match score with a threshold of 0.8 to identify the relevant tables based on their captions. Note that not all instances have these formats (see Table 4) due to LaTeXML conversion errors, low fuzzy match score, discrepancies between captions in the gold data and LaTeX files or a scholarly paper not being available on arXiv anymore. We also exclude cases with multiple tables sharing the same caption but annotated separately, as it is challenging to accurately link the corresponding HTML/XML code for each table. HTML in LogicNLG and Logic2Text are retrieved from the Wikipedia pages. However, due to the lack of metadata on the data collection timestamps, we choose a time interval close to the year of publication of these datasets for our search in the Wikipedia archive. To extract the relevant tables, we employ a cosine similarity comparison against the gold tables, using a threshold of 0.9. Since Wikipedia is constantly updated, we further manually check the results and filter out cases where the mismatch affects the ground truth, e. g., cell values being out of date or the removal/addition of both rows and columns. Note that for all subsets except SciGen, we follow the PMC table formatting rules[19] to obtain XML. Additionally, all generated HTML underwent automatic validation using the PyTidyLib[20] package.

LaTeX. Similar to HTML/XML, we obtain LaTeX from the source scholarly papers in SciGen (see §2.2) and extract the target tables based on their captions using the fuzzy match. Some instances are excluded due to low similarity scores (below 0.8), parsing errors or lack of LaTeX source code (tables from ACL papers). For numericNLG and PubTables-1M tables, LaTeX is generated from HTML. This process involves preprocessing the HTML code to replace symbols, such as Greek letters and mathematical operators, with their LaTeX equivalents. The resulting HTML is then converted to a dataframe and subsequently to LaTeX using pandas.

Dict. The conventions of already available linearised tables in SciGen, numericNLG, LogicNLG, and Logic2Text are slightly diverse. In particular, the distinction between column and row heads exists only in numericNLG. Furthermore, compared to LogicNLG and Logic2Text, header hierarchy is preserved in numericNLG and SciGen by merging headers and subheaders into a single string.

[19] https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/dobs.html#dob-tables

[20] https://countergram.github.io/pytidylib/

D Image aspect ratios

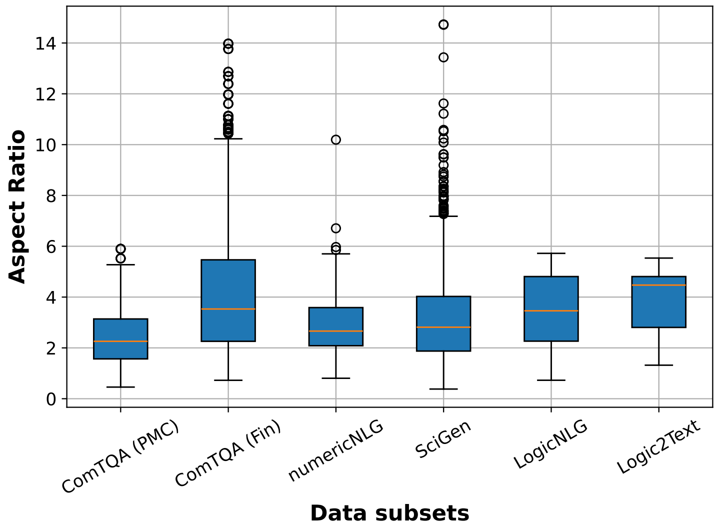

Figure 12: Distribution of image aspect ratios (width/height) across subsets in the TableEval benchmark. Each box represents the interquartile range (IQR), with the central orange line indicating the median. Circles denote outliers, while whiskers (set to 1.5 × IQR by default) extend to the minimum and maximum non-outlier values. Here “Fin” stands for FinTabNet, while “PMC” denotes PubTables-1M.

E Prompts

ComTQA (FinTabNet):

```
Refer to the provided table and answer the question. Question: {question}
```

ComTQA (PubTables-1M):

```
Refer to the provided table and answer the question. Question: {question}.
Table caption: {caption}. Table footnote: {footnote}.
```

SciGen:

```
Describe the given table focusing on the most important findings reported by
reasoning over its content. The summary must be factual, coherent, and
well-written. Do not introduce new information or speculate. Table caption:
{caption}
```

numericNLG:

```
Describe the given table focusing on the insights and trends revealed by the
results. The summary must be factual, coherent, and well-written. Do not
introduce new information or speculate. Table caption: {caption}
```

Logic2Text:

```
Generate a one sentence statement based on the table and logical form. Logical
form: {logical_form}. Table title: {title}
```

LogicNLG:

```
Based on a given table, fill in the entities masked by [ENT] in the following
sentence: {sentence}. Output the sentence with filled in masked entities.
Table title: {title}
```

Figure 13: Prompts used for experiments based on images of tables.

ComTQA (FinTabNet):

`Refer to the provided table and answer the question. Question: {question}. Table: {table}.`

ComTQA (PubTables-1M):

`Refer to the provided table and answer the question. Question: {question}. Table: {table}.`

SciGen:

`Describe the given table focusing on the most important findings reported by reasoning over its content. The summary must be factual, coherent, and well-written. Do not introduce new information or speculate. Table: {table}.`

numericNLG:

`Describe the given table focusing on the insights and trends revealed by the results. The summary must be factual, coherent, and well-written. Do not introduce new information or speculate. Table: {table}.`

Logic2Text:

`Generate a one sentence statement based on the table and logical form. Logical form: {logical_form}. Table title: {title}. Table: {table}.`

LogicNLG:

`Based on a given table, fill in the entities masked by [ENT] in the following sentence: {sentence}. Output the sentence with filled in masked entities. Table title: {title}. Table: {table}.`

Figure 14: Prompts used for experiments based on textual representations of tables.

F Experimental results

Metric	Dict	HTML	Image	LaTeX	XML
BertScore.F1	0.85	0.84	**0.86**	0.84	0.85
BLEU-1	0.16	0.15	**0.19**	0.16	0.16
BLEU-2	0.09	0.09	**0.12**	0.09	0.09
BLEU-3	0.06	0.06	**0.09**	0.06	0.07
BLEU-4	0.04	0.04	**0.06**	0.05	0.05
BLEURT	−0.51	−0.55	**−0.42**	−0.54	−0.53
METEOR	0.24	0.24	**0.25**	0.24	0.24
MoverScore	0.54	0.53	**0.56**	0.54	0.54
ROUGE-1.F1	0.30	0.29	**0.38**	0.29	0.29
ROUGE-2.F1	0.15	0.14	**0.20**	0.15	0.15
ROUGE-3.F1	0.09	0.09	**0.12**	0.09	0.09
ROUGE-4.F1	0.06	0.06	**0.08**	0.07	0.06
ROUGE-L.F1	0.24	0.23	**0.32**	0.24	0.24
SacreBLEU	0.04	0.04	**0.08**	0.05	0.05

Table 5: Values across evaluation metrics for table formats averaged over data subsets and models.

Metric	Dict	HTML	Image	LaTeX	XML
BertScore.F1	0.83	0.84	**0.86**	0.83	0.84
BLEU-1	0.02	0.02	**0.05**	0.02	0.02
BLEU-2	0.01	0.01	**0.03**	0.01	0.01
BLEU-3	0.01	0.01	**0.02**	0.01	0.01
BLEU-4	0.01	0.01	**0.02**	0.01	0.01
BLEURT	−0.58	−0.55	**−0.39**	−0.59	−0.54
METEOR	0.06	0.07	**0.08**	0.06	0.07
MoverScore	0.50	0.50	**0.53**	0.49	0.50
ROUGE-1.F1	0.14	0.14	**0.27**	0.14	0.15
ROUGE-2.F1	0.08	0.08	**0.17**	0.08	0.09
ROUGE-3.F1	0.03	0.03	**0.05**	0.03	0.03
ROUGE-4.F1	0.01	0.01	**0.02**	0.01	0.01
ROUGE-L.F1	0.13	0.14	**0.27**	0.14	0.15
SacreBLEU	0.01	0.02	**0.04**	0.01	0.02

Table 6: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for ComTQA (FinTabNet) subset for individual formats averaged over models.

Metric	Dict	HTML	Image	LaTeX	XML
BertScore.F1	0.82	0.82	**0.85**	0.82	0.82
BLEU-1	0.03	0.03	**0.05**	0.03	0.03
BLEU-2	0.02	0.02	**0.03**	0.02	0.02
BLEU-3	0.01	**0.02**	**0.02**	0.01	**0.02**
BLEU-4	0.01	0.01	**0.02**	0.01	0.01
BLEURT	−0.73	−0.72	**−0.59**	−0.73	−0.72
METEOR	**0.09**	0.10	**0.09**	**0.09**	0.10
MoverScore	0.48	0.48	**0.51**	0.48	0.48
ROUGE-1.F1	0.12	0.12	**0.22**	0.12	0.12
ROUGE-2.F1	0.06	0.06	**0.11**	0.06	0.06
ROUGE-3.F1	0.03	0.03	**0.04**	0.03	0.03
ROUGE-4.F1	0.02	0.02	**0.03**	0.02	0.02
ROUGE-L.F1	0.12	0.12	**0.22**	0.11	0.12
SacreBLEU	0.01	0.01	**0.04**	0.01	0.01

Table 7: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for ComTQA (PubTables-1M) subset for individual formats averaged over models.

Metric	Dict	HTML	Image	LaTeX	XML
BertScore.F1	0.88	0.88	**0.89**	0.88	0.88
BLEU-1	**0.24**	**0.24**	0.22	**0.24**	**0.24**
BLEU-2	**0.13**	**0.13**	0.12	**0.13**	**0.13**
BLEU-3	0.07	0.07	0.07	0.07	**0.08**
BLEU-4	0.04	0.04	0.04	0.04	**0.05**
BLEURT	−0.14	−0.11	−0.19	**−0.09**	**−0.09**
METEOR	0.35	0.37	0.33	0.37	**0.38**
MoverScore	0.59	**0.60**	**0.60**	**0.60**	**0.60**
ROUGE-1.F1	0.48	**0.49**	**0.49**	**0.49**	**0.49**
ROUGE-2.F1	0.23	0.24	0.24	**0.25**	0.24
ROUGE-3.F1	0.12	0.13	0.12	**0.14**	0.13
ROUGE-4.F1	0.06	0.07	0.07	**0.08**	0.07
ROUGE-L.F1	0.37	**0.39**	**0.39**	0.38	0.38
SacreBLEU	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**

Table 8: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for Logic2Text subset for individual formats averaged over models.

Metric	Dict	HTML	Image	LaTeX	XML
BertScore.F1	0.87	0.88	**0.91**	0.89	0.88
BLEU-1	0.32	0.33	**0.51**	0.36	0.36
BLEU-2	0.26	0.27	**0.43**	0.30	0.29
BLEU-3	0.21	0.23	**0.35**	0.25	0.24
BLEU-4	0.17	0.18	**0.28**	0.20	0.20
BLEURT	−0.46	−0.47	**−0.13**	−0.40	−0.41
METEOR	0.52	0.53	**0.63**	0.55	0.55
MoverScore	0.60	0.59	**0.64**	0.61	0.60
ROUGE-1.F1	0.48	0.48	**0.69**	0.52	0.51
ROUGE-2.F1	0.38	0.38	**0.55**	0.41	0.40
ROUGE-3.F1	0.31	0.30	**0.45**	0.34	0.33
ROUGE-4.F1	0.25	0.25	**0.37**	0.28	0.27
ROUGE-L.F1	0.46	0.47	**0.67**	0.51	0.49
SacreBLEU	0.13	0.15	**0.28**	0.16	0.16

Table 9: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for LogicNLG subset for individual formats averaged over models.

Metric	Dict	HTML	Image	LaTeX	XML
BertScore.F1	0.83	**0.84**	0.83	**0.84**	**0.84**
BLEU-1	0.16	**0.18**	0.16	**0.18**	**0.18**
BLEU-2	0.06	**0.07**	**0.07**	**0.07**	**0.07**
BLEU-3	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**
BLEU-4	0.01	**0.02**	0.01	**0.02**	**0.02**
BLEURT	−0.58	−0.54	−0.60	−0.54	**−0.53**
METEOR	0.19	**0.21**	0.19	**0.21**	**0.21**
MoverScore	0.52	**0.53**	**0.53**	**0.53**	**0.53**
ROUGE-1.F1	0.28	0.31	0.30	**0.32**	**0.32**
ROUGE-2.F1	0.06	**0.08**	0.07	**0.08**	**0.08**
ROUGE-3.F1	0.02	0.02	0.02	0.02	**0.03**
ROUGE-4.F1	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**
ROUGE-L.F1	0.16	**0.17**	**0.17**	**0.17**	**0.17**
SacreBLEU	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**

Table 10: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for numericNLG subset for individual formats averaged over models.

Metric	Dict	HTML	Image	LaTeX	XML
BertScore.F1	**0.84**	0.81	**0.84**	0.81	0.81
BLEU-1	**0.16**	0.11	0.15	0.11	0.11
BLEU-2	**0.07**	0.03	**0.07**	0.03	0.03
BLEU-3	**0.03**	0.01	**0.03**	0.01	0.01
BLEU-4	**0.02**	0.00	**0.02**	0.00	0.00
BLEURT	**−0.59**	−0.90	−0.64	−0.91	−0.90
METEOR	**0.20**	0.13	0.19	0.13	0.13
MoverScore	**0.53**	0.50	**0.53**	0.50	0.50
ROUGE-1.F1	**0.30**	0.18	0.29	0.18	0.18
ROUGE-2.F1	**0.07**	0.02	**0.07**	0.02	0.02
ROUGE-3.F1	0.02	0.00	**0.03**	0.00	0.00
ROUGE-4.F1	**0.01**	0.00	**0.01**	0.00	0.00
ROUGE-L.F1	**0.17**	0.11	**0.17**	0.11	0.11
SacreBLEU	**0.03**	0.01	**0.03**	0.01	0.01

Table 11: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for SciGen subset for individual formats averaged over models.

Metric	Dict	HTML	LaTeX	XML
BertScore.F1	**0.83**	**0.83**	**0.83**	**0.83**
BLEU-1	**0.12**	**0.12**	0.11	0.11
BLEU-2	**0.06**	**0.06**	**0.06**	**0.06**
BLEU-3	0.03	**0.04**	**0.04**	**0.04**
BLEU-4	**0.02**	**0.02**	**0.02**	**0.02**
BLEURT	**−0.64**	−0.67	−0.67	−0.66
METEOR	0.20	**0.21**	0.20	**0.21**
MoverScore	**0.52**	**0.52**	**0.52**	**0.52**
ROUGE-1.F1	**0.23**	**0.23**	**0.23**	**0.23**
ROUGE-2.F1	0.09	**0.10**	**0.10**	**0.10**
ROUGE-3.F1	**0.05**	**0.05**	**0.05**	**0.05**
ROUGE-4.F1	**0.03**	**0.03**	**0.03**	**0.03**
ROUGE-L.F1	0.17	**0.18**	**0.18**	**0.18**
SacreBLEU	**0.02**	**0.02**	**0.02**	**0.02**

Table 12: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for Llama-3.2-3B-Instruct and individual text formats averaged over data subsets.

Metric	Dict	HTML	LaTeX	XML
BertScore.F1	**0.85**	**0.85**	**0.85**	**0.85**
BLEU-1	0.17	0.15	**0.18**	0.17
BLEU-2	0.10	0.09	**0.11**	0.10
BLEU-3	0.06	0.06	**0.07**	**0.07**
BLEU-4	0.04	0.04	**0.05**	**0.05**
BLEURT	**−0.48**	−0.54	**−0.48**	−0.49
METEOR	**0.25**	0.24	**0.25**	**0.25**
MoverScore	**0.54**	**0.54**	**0.54**	**0.54**
ROUGE-1.F1	0.33	0.31	**0.34**	0.33
ROUGE-2.F1	0.17	0.16	**0.18**	**0.18**
ROUGE-3.F1	**0.11**	0.10	**0.11**	**0.11**
ROUGE-4.F1	0.07	0.07	**0.08**	**0.08**
ROUGE-L.F1	0.27	0.26	**0.28**	**0.28**
SacreBLEU	0.04	0.04	**0.05**	**0.05**

Table 13: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for Mistral-Nemo-Instruct-2407 and individual text formats averaged over data subsets.

Metric	**Dict**	**HTML**	**LaTeX**	**XML**
BertScore.F1	**0.84**	**0.84**	**0.84**	**0.84**
BLEU-1	**0.13**	**0.13**	**0.13**	**0.13**
BLEU-2	**0.07**	**0.07**	**0.07**	**0.07**
BLEU-3	0.04	**0.05**	**0.05**	**0.05**
BLEU-4	**0.03**	**0.03**	**0.03**	**0.03**
BLEURT	**−0.54**	−0.55	−0.57	−0.56
METEOR	0.23	**0.24**	0.23	**0.24**
MoverScore	**0.53**	**0.53**	**0.53**	**0.53**
ROUGE-1.F1	**0.26**	**0.26**	**0.26**	**0.26**
ROUGE-2.F1	0.12	**0.13**	0.12	**0.13**
ROUGE-3.F1	**0.07**	**0.07**	**0.07**	**0.07**
ROUGE-4.F1	**0.05**	**0.05**	**0.05**	**0.05**
ROUGE-L.F1	0.20	**0.21**	0.20	0.20
SacreBLEU	**0.03**	**0.03**	**0.03**	**0.03**

Table 14: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for Qwen2.5-14B-Instruct and individual text formats averaged over data subsets.

Metric	**Dict**	**HTML**	**LaTeX**	**XML**
BertScore.F1	**0.86**	**0.86**	**0.86**	**0.86**
BLEU-1	0.21	**0.22**	0.21	**0.22**
BLEU-2	0.13	0.14	0.14	**0.15**
BLEU-3	0.10	**0.11**	0.10	**0.11**
BLEU-4	0.08	**0.09**	0.08	**0.09**
BLEURT	**−0.37**	−0.39	−0.41	−0.38
METEOR	0.26	**0.27**	0.26	**0.27**
MoverScore	**0.56**	**0.56**	0.55	**0.56**
ROUGE-1.F1	**0.38**	0.37	0.36	0.37
ROUGE-2.F1	**0.21**	**0.21**	0.20	**0.21**
ROUGE-3.F1	0.13	**0.14**	0.13	**0.14**
ROUGE-4.F1	**0.10**	**0.10**	**0.10**	**0.10**
ROUGE-L.F1	**0.32**	0.31	0.30	0.31
SacreBLEU	0.09	0.10	0.10	**0.11**

Table 16: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for Gemini-2.0-Flash and individual text formats averaged over data subsets.

Metric	**Dict**	**HTML**	**LaTeX**	**XML**
BertScore.F1	**0.84**	**0.84**	**0.84**	**0.84**
BLEU-1	**0.16**	0.15	**0.16**	0.15
BLEU-2	**0.09**	0.08	**0.09**	**0.09**
BLEU-3	0.06	0.06	**0.07**	0.06
BLEU-4	0.04	0.04	**0.05**	**0.05**
BLEURT	**−0.54**	−0.59	−0.57	−0.57
METEOR	**0.24**	0.23	**0.24**	0.23
MoverScore	**0.53**	**0.53**	**0.53**	**0.53**
ROUGE-1.F1	**0.28**	0.27	**0.28**	**0.28**
ROUGE-2.F1	0.13	0.13	**0.14**	0.13
ROUGE-3.F1	0.08	0.08	**0.09**	0.08
ROUGE-4.F1	**0.06**	0.05	**0.06**	**0.06**
ROUGE-L.F1	0.22	0.21	**0.23**	0.22
SacreBLEU	0.03	0.03	**0.04**	0.03

Table 15: Raw values of BertScore.F1, BLEU-N.F1, BLEURT, METEOR, MoverScore, ROUGE-N.F1, ROUGE-L.F1, and SacreBLEU for Qwen2.5-3B-Instruct and individual text formats averaged over data subsets.

Metric	**Non-Scientific**	**Scientific**
BertScore.F1	**0.87**	0.83
BLEU-1	**0.21**	0.11
BLEU-2	**0.15**	0.04
BLEU-3	**0.11**	0.02
BLEU-4	**0.09**	0.01
BLEURT	**−0.34**	−0.68
METEOR	**0.33**	0.15
MoverScore	**0.57**	0.51
ROUGE-1.F1	**0.40**	0.22
ROUGE-2.F1	**0.25**	0.06
ROUGE-3.F1	**0.17**	0.02
ROUGE-4.F1	**0.12**	0.01
ROUGE-L.F1	**0.36**	0.15
SacreBLEU	**0.08**	0.02

Table 17: Values across evaluation metrics for scientific and non-scientific domains averaged over data subsets, models, and table formats.

Metric	ComTQA (FinTabNet)	ComTQA (PubTables-1M)	Logic2Text	LogicNLG	numericNLG	SciGen
BertScore.F1	0.84	0.83	0.88	**0.89**	0.83	0.82
BLEU-1	0.03	0.04	0.23	**0.38**	0.17	0.13
BLEU-2	0.02	0.02	0.13	**0.31**	0.07	0.04
BLEU-3	0.01	0.02	0.07	**0.26**	0.03	0.02
BLEU-4	0.01	0.01	0.04	**0.20**	0.01	0.01
BLEURT	−0.53	−0.70	**−0.13**	−0.37	−0.56	−0.79
METEOR	0.07	0.09	0.36	**0.56**	0.20	0.16
MoverScore	0.50	0.49	0.60	**0.61**	0.53	0.51
ROUGE-1.F1	0.17	0.14	0.49	**0.54**	0.31	0.23
ROUGE-2.F1	0.10	0.07	0.24	**0.42**	0.07	0.04
ROUGE-3.F1	0.03	0.03	0.13	**0.34**	0.02	0.01
ROUGE-4.F1	0.01	0.02	0.07	**0.28**	0.01	0.00
ROUGE-L.F1	0.17	0.14	0.38	**0.52**	0.17	0.13
SacreBLEU	0.02	0.02	0.05	**0.18**	0.03	0.02

Table 18: Values across evaluation metrics for each data subset averaged over models and table formats.

Model	Bert-Score.F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEURT	METEOR	Mover-Score	ROUGE-1.F1	ROUGE-2.F1	ROUGE-3.F1	ROUGE-4.F1	ROUGE-L.F1	Sacre-BLEU
Baseline														
Gemini-2.0-Flash_mm	**0.87**	**0.22**	**0.14**	**0.11**	**0.08**	**−0.35**	**0.27**	**0.56**	**0.40**	**0.22**	**0.14**	**0.10**	**0.33**	**0.11**
Gemini-2.0-Flash_llm	0.86	0.21	**0.14**	**0.11**	**0.08**	−0.39	0.26	**0.56**	0.37	0.20	**0.14**	**0.10**	0.31	0.10
MLLMs														
Idefics3-8B-Llama3	**0.88**	**0.19**	**0.12**	**0.09**	**0.07**	**−0.36**	0.23	**0.59**	**0.47**	**0.27**	**0.13**	**0.09**	**0.42**	**0.11**
Qwen2.5-VL-3B-Instruct	0.85	0.18	**0.12**	**0.09**	**0.07**	−0.51	0.25	0.55	0.34	0.18	0.11	0.08	0.28	0.07
Qwen2.5-VL-7B-Instruct	0.86	**0.19**	**0.12**	0.08	0.06	−0.39	**0.27**	0.55	0.36	0.19	0.12	**0.09**	0.30	0.07
llama3-llava-next-8b-hf	0.85	0.16	0.10	0.06	0.04	−0.50	0.24	0.54	0.31	0.15	0.09	0.06	0.25	0.04
LLMs														
Mistral-Nemo-Instruct-2407	**0.85**	**0.17**	**0.10**	**0.07**	**0.05**	**−0.50**	**0.25**	**0.54**	**0.33**	**0.17**	**0.11**	**0.07**	**0.27**	**0.04**
Qwen2.5-3B-Instruct	0.84	0.15	0.09	0.06	0.04	−0.57	0.24	0.53	0.28	0.13	0.08	0.06	0.22	0.03
Qwen2.5-14B-Instruct	0.84	0.13	0.07	0.05	0.03	−0.56	0.24	0.53	0.26	0.12	0.07	0.05	0.20	0.03
Llama-3.2-3B-Instruct	0.83	0.12	0.06	0.04	0.02	−0.66	0.20	0.52	0.23	0.10	0.05	0.03	0.18	0.02

Table 19: Values across evaluation metrics for individual models averaged over data subsets and table formats.

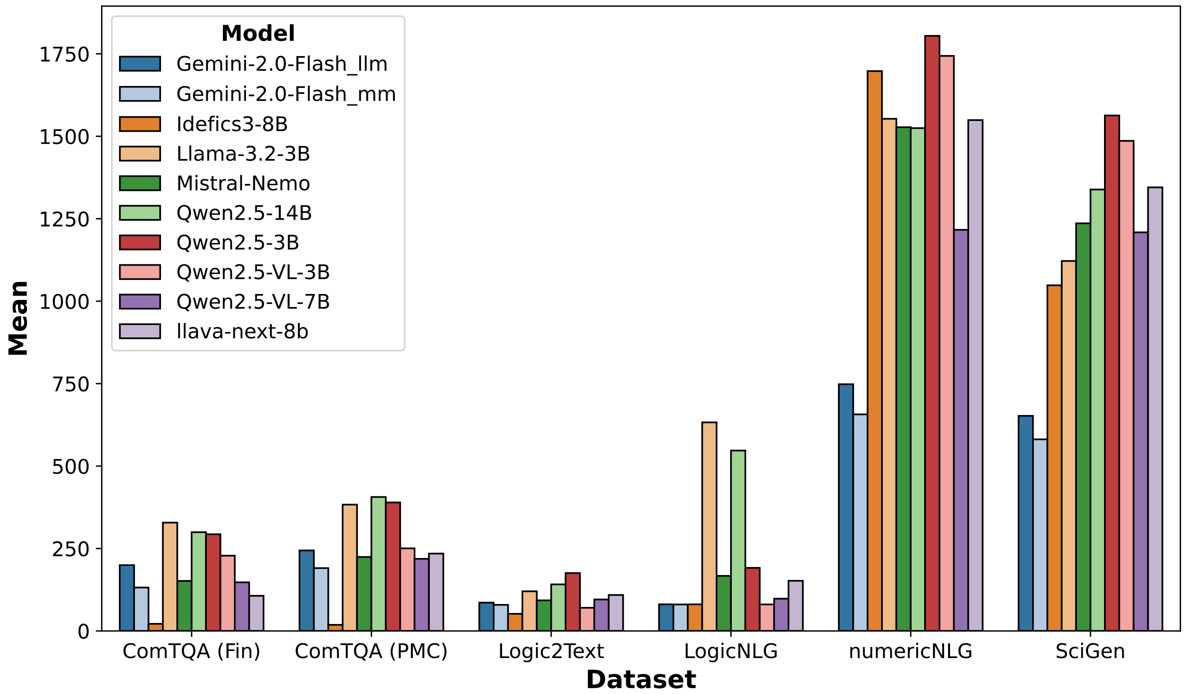

Figure 15: Mean prediction lengths (in characters) for each model and data subset. Here "_llm" and "_mm" are used to distinguish between text and image input for Gemini-2.0-Flash, respectively.

G Case Study

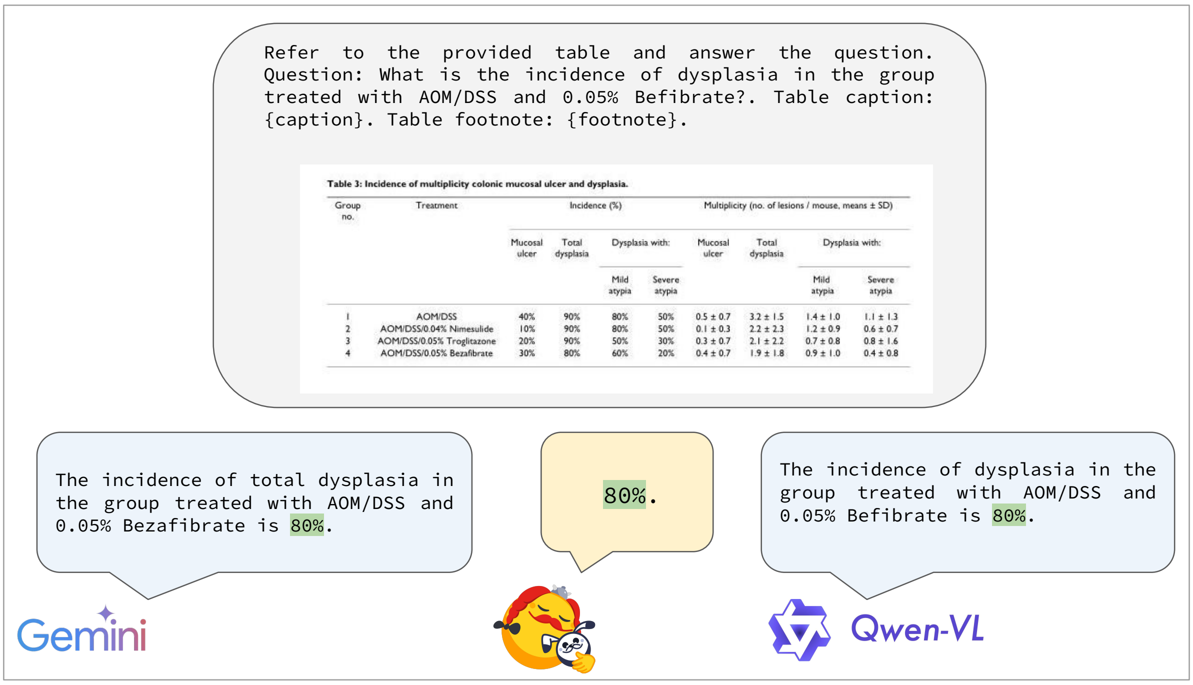

Table 3: Incidence of multiplicity colonic mucosal ulcer and dysplasia.

Group no.	Treatment	Incidence (%)				Multiplicity (no. of lesions / mouse, means ± SD)			
		Mucosal ulcer	Total dysplasia	Dysplasia with:		Mucosal ulcer	Total dysplasia	Dysplasia with:	
				Mild atypia	Severe atypia			Mild atypia	Severe atypia
1	AOM/DSS	40%	90%	80%	50%	0.5 ± 0.7	3.2 ± 1.5	1.4 ± 1.0	1.1 ± 1.3
2	AOM/DSS/0.04% Nimesulide	10%	90%	80%	50%	0.1 ± 0.3	2.2 ± 2.3	1.2 ± 0.9	0.6 ± 0.7
3	AOM/DSS/0.05% Troglitazone	20%	90%	50%	30%	0.3 ± 0.7	2.1 ± 2.2	0.7 ± 0.8	0.8 ± 1.6
4	AOM/DSS/0.05% Bezafibrate	30%	80%	60%	20%	0.4 ± 0.7	1.9 ± 1.8	0.9 ± 1.0	0.4 ± 0.8

Figure 16: An example illustrating differences in prediction length across Idefics3, Gemini-2.0-Flash, and Qwen2.0-VL (7B) models on a sample from the ComTQA (PubTables-1M) subset.

H Additional interpretability analyses

Mistral-Nemo vs. Llama3. The following figures show further examples of feature attribution and log-probability analysis comparing Mistral-Nemo with Llama3.

In Figure 17 (ComTQA FinTabNet), Mistral-Nemo correctly predicts the answer, while Llama3 fails. We find a key difference in the attribution pattern around the columns *"2014"* and *"2013"*, where Mistral-Nemo assigns a slightly higher score (lighter blue) than Llama3. In the log-probability analysis, we see high uncertainty in Llama3 generating the final answer starting with *"1"*. On the contrary, Mistral-Nemo shows a high level of confidence in the predicted value.

In Figure 18 (ComTQA PubTables-1M), both models generate incorrect answers. For Mistral-Nemo, one can barely see any attribution in the decisive row of the table. For Llama3, there is a slightly higher attribution for *"Beer"* in *"Lung-Beer"*. We also observe that the tokeniser splits the number into *"496"* and *"6"*. A plausible explanation for the failure is that when it processes *"Lung Stanford"* with 918 genes, it likely finds it to be higher than 496 (ignoring the fourth digit *"6"*). Regarding the log-probabilities, the decision of which feature to name after *"the most number of genes is"* is controversial for both models, judged by the low confidence in the following token.

In Figure 19 (ComTQA PubTables-1M), Mistral-Nemo solves the task correctly, whereas Llama3 fails to distinguish *"VRP-HA"* from *"VRP-neu"* and is not confident in the predicted value (*10*). Mistral-Nemo focuses on the *"VRP-HA"* row in the table more than the similar alternative *"VRP-neu"* and generally finds the relevant feature name in the question to be more important, judging by the attribution patterns. When we compare this to the log-probabilities, the model is very confident about its decision (*"VRP-HA"*) throughout the generation.

Dict vs. LaTeX input format. The following figures show examples of feature attribution and log-probability analysis. We compare predictions across Dict vs. LaTeX representations of tables for Mistral-Nemo and Llama3 based on instances from the LogicNLG subset.

In Figure 20, Mistral-Nemo correctly predicts the missing entities with a high level of confidence. We notice high similarity between the input attribution patterns across two formats. In both cases, one of the most relevant tokens (month *"August"*) is correctly identified to produce the right answer according to the ground truth and hence receives high attribution. The model focuses on the tokens relevant to the task and does not pay much attention to LaTeX formatting tags, since the respective tokens generally remain barely considered throughout the generation. However, we can see some decreases in model confidence at the end of the generation (*"games before"*).

In Figure 21, Llama3 generates the wrong responses in both cases. However, the Dict variant also makes the model focus on bracketing, separators, and punctuation quite often. Only for LaTeX, there is a noticeably lower confidence about generating *"Electra"* as the play of choice. For both representations of the table, however, Llama3 is not certain about the last two entities (*"Cyprus and Romania"*, *"Cyprus and Greece"*), which are either fully or partially incorrect according to the ground truth (*"Greece and Italy"*).

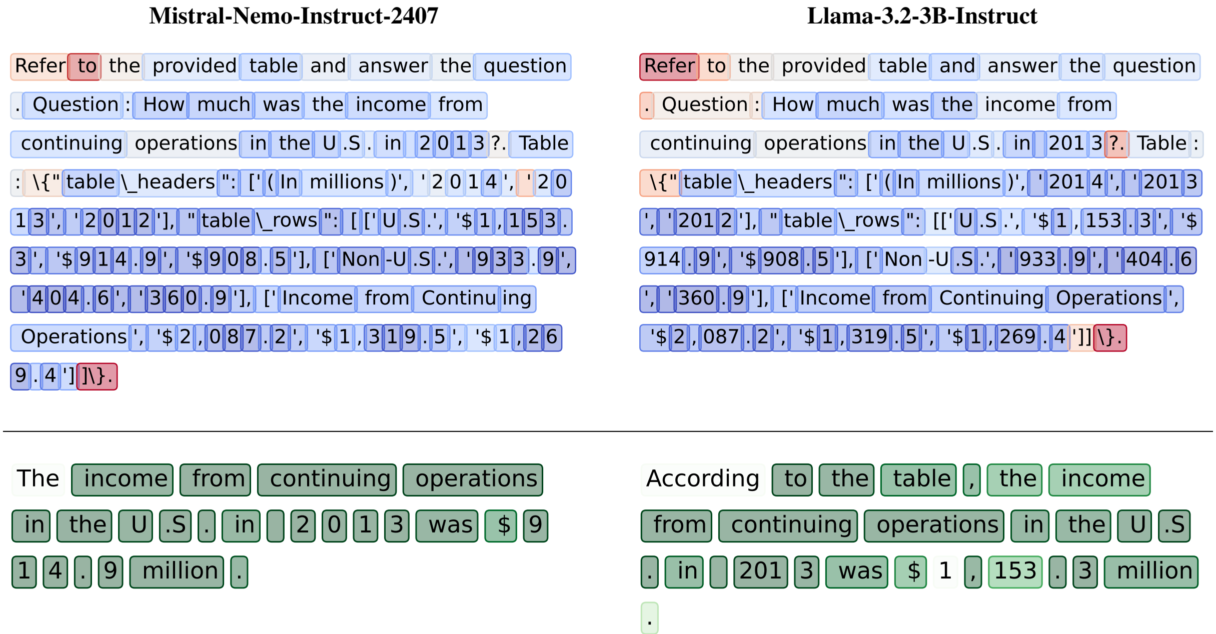

Figure 17: Interpretability analysis for the ComTQA (FinTabNet) instance with a table represented in a Dict format. The ground truth is *“$914.9 million”*. The visualisation follows the same procedure as Figure 5.

Mistral-Nemo-Instruct-2407

Refer to the provided table and answer the question. Question: Which dataset has the most number of genes?. Table: \{"table_title": Table 5, "table_caption": Random data simulations of real data sets. This table compares the results found from the real data (Real column) to two different types of random data. The Random column contains the experimentally determined largest number of pairs found from 10 simulation runs using a random data matrix (drawn from a uniform distribution) where the number of genes and class sizes is the same as the indicated for the real data. The Label Shuffled column contains the experimentally determined largest number of pairs found from 30 simulation runs where the class labels were randomly shuffled. In the samples column, the number in parenthesis is the number of positive samples. The numbers after the slash are the number of single genes found. Label shuffling leads to more pairs found "by chance" only for the smaller data sets. The small data sets have large numbers of pairs expected "by chance".", "table_headers": ['Data set', 'Samples', 'Genes', 'Real', 'Random', 'Label Shuffled'], "table_subheaders": [], "table_rows": [['GIST', '19(6)', '1987', '137981/74', '2706/0', '4622/2'], ['BreastBRCA(brca1 vs brca2)', '15(7)', '3226', '143574/18', '20563/2', '53900/11'], ['BreastBRCA(brca1 \& brca2 vs Sporadic)', '22(7)', '3226', '2114/0', '1286/1', '0/0'], ['Cutaneous', '38(7)', '3613', '596/0', '62/0', '24/0'], ['LungStanford', '52(13)', '918', '486/2', '0/0', '0/0'], ['LungBeer', '96(10)', '4966', '22102/5', '0/0', '0/0'], ['Prostate', '34(9)', '3958', '249662/52', '57/0', '13/0']], "table_footnote": None\}.

Based on the provided table, the dataset with the most number of genes is "Prostate" with 3958 genes.

Llama-3.2-3B-Instruct

Refer to the provided table and answer the question. Question: Which dataset has the most number of genes?. Table: \{"table_title": Table 5, "table_caption": Random data simulations of real data sets. This table compares the results found from the real data (Real column) to two different types of random data. The Random column contains the experimentally determined largest number of pairs found from 10 simulation runs using a random data matrix (drawn from a uniform distribution) where the number of genes and class sizes is the same as the indicated for the real data. The Label Shuffled column contains the experimentally determined largest number of pairs found from 30 simulation runs where the class labels were randomly shuffled. In the samples column, the number in parenthesis is the number of positive samples. The numbers after the slash are the number of single genes found. Label shuffling leads to more pairs found "by chance" only for the smaller data sets. The small data sets have large numbers of pairs expected "by chance".", "table_headers": ['Data set', 'Samples', 'Genes', 'Real', 'Random', 'Label Shuffled'], "table_subheaders": [], "table_rows": [['GIST', '19(6)', '1987', '137981/74', '2706/0', '4622/2'], ['BreastBRCA(brca1 vs brca2)', '15(7)', '3226', '143574/18', '20563/2', '53900/11'], ['BreastBRCA(brca1 \& brca2 vs Sporadic)', '22(7)', '3226', '2114/0', '1286/1', '0/0'], ['Cutaneous', '38(7)', '3613', '596/0', '62/0', '24/0'], ['LungStanford', '52(13)', '918', '486/2', '0/0', '0/0'], ['LungBeer', '96(10)', '4966', '22102/5', '0/0', '0/0'], ['Prostate', '34(9)', '3958', '249662/52', '57/0', '13/0']], "table_footnote": None\}.

Based on the table, the dataset with the most number of genes is 'LungStanford' with 918 genes.

Figure 18: Interpretability analysis for the ComTQA (PubTables-1M) instance with a table represented in a Dict format. The ground truth is *"LungBeer"*. The visualisation follows the same procedure as Figure 5.

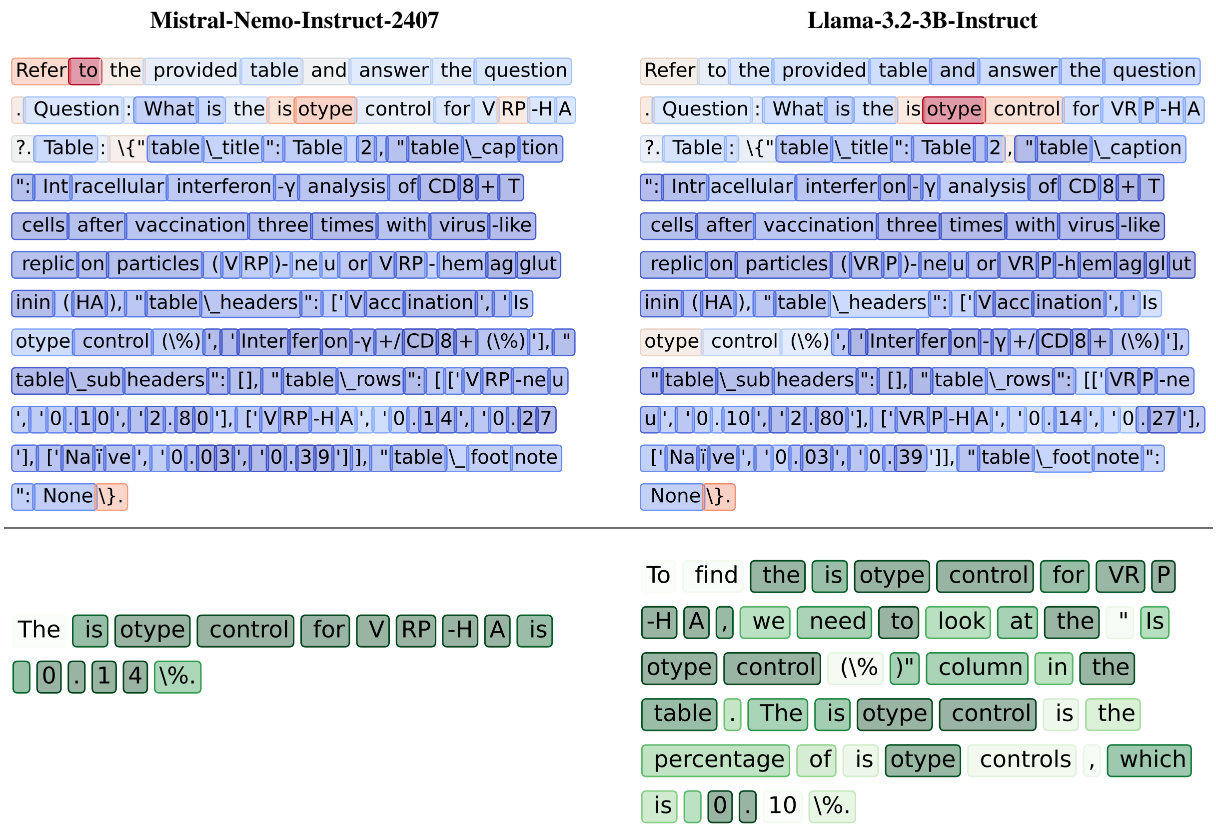

Figure 19: Interpretability analysis for ComTQA (PubTables-1M) instance with the Dict format. The ground truth is *"0.14%"*. The visualisation follows the same procedure as Figure 5.

Figure 20: Interpretability analysis the LogicNLG instance comparing the Dict (left) with the LaTeX (right) input format of the table. The ground truth is *"all game in the table for the 1908 Vfl Season were played on 15 August 1908 with no game after or before"*. The visualisation follows the same procedure as Figure 5.

Figure 21: Interpretability analysis for the LogicNLG instance comparing the Dict (left) with the LaTeX (right) input format of the table. The ground truth is *"the play Medea was performed in Greece and Italy"*. The visualisation follows the same procedure as Figure 5.

	Years Ended December 31,						
	2014	2013	2012	2014 vs. 2013		2013 vs. 2012	
				$ Change	% Change	$ Change	% Change
Routing	$ 2,223.9	$ 2,318.0	$ 2,037.6	$ (94.1)	(4)%	$ 280.4	14%
Switching	721.2	638.0	554.8	83.2	13 %	83.2	15%
Security	463.6	563.9	669.7	(100.3)	(18)%	(105.8)	(16)%
Total Product	3,408.7	3,519.9	3,262.1	(111.2)	(3)%	257.8	8%
Percentage of net revenues	*73.7 %*	*75.4 %*	*74.7 %*				
Total Service	1,218.4	1,149.2	1,103.3	69.2	6 %	45.9	4%
Percentage of net revenues	*26.3 %*	*24.6 %*	*25.3 %*				
Total net revenues	$ 4,627.1	$ 4,669.1	$ 4,365.4	$ (42.0)	(1)%	$ 303.7	7%

Figure 22: Table image corresponding to the ComTQA (FinTabNet) example in Figure 5.