PaperUnPlot: Benchmarking Chart-to-Table in the Wild

Daniele Bertillo
Roma Tre University
Rome, Italy
daniele.bertillo@uniroma3.it

Giorgio Melchiorri
Roma Tre University
Rome, Italy
gio.melchiorri@stud.uniroma3.it

Paolo Merialdo
Roma Tre University
Rome, Italy
paolo.merialdo@uniroma3.it

ABSTRACT

Charts are essential for communicating quantitative information in scientific literature. Yet automated extraction of the underlying data, the chart-to-table task, remains an open challenge. To assess how well current benchmarks reflect real-world scientific figures, we conduct a large-scale analysis of chart type distributions across PubMed Central and arXiv, covering over 356,000 chart images. Our findings reveal a significant discrepancy between the charts prevalent in real-world publications and those covered by existing benchmarks, exposing a systematic blind spot in current evaluation frameworks. To address this, we introduce PaperUnPlot, a benchmark for the chart-to-table task comprising 950 manually annotated real chart-table pairs across 10 chart categories, including those neglected by prior work, drawn from PMC and arXiv, and complemented by a paired synthetic split for controlled evaluation. We also propose Normalized Mapping Similarity (NMS), an extension of the standard RMS metric that handles narrow and logarithmic axes common in scientific charts. We evaluate eight models against our benchmark, including GPT-4o, GPT-4o Mini, Gemini 2.5 Flash, Claude Sonnet 4.5, three small open-weight models, and the task-specific model DePlot. Our results show that small open-weight models remain far from competitive with the best commercial models, highlighting an open challenge for resource-constrained deployment. Performance also varies widely across chart types, and all models score consistently higher on synthetic than on real charts, revealing that existing benchmarks systematically overestimate practical capabilities.

VLDB Workshop Reference Format:
Daniele Bertillo, Giorgio Melchiorri, and Paolo Merialdo. PaperUnPlot: Benchmarking Chart-to-Table in the Wild. VLDB 2026 Workshop: Tabular Data Analysis (TaDA).

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/giorgiomelch/about_chart_to_table.

1 INTRODUCTION

Charts are a key tool for communicating quantitative findings in scientific literature, appearing across every discipline. Automatically extracting the underlying data from chart images, a task known as chart-to-table, would unlock a wide range of downstream applications, including document understanding, scientific data mining, and accessibility tools for visually impaired users. For example, consider a researcher surveying hundreds of papers to compare model performance across benchmarks, or a financial analyst extracting trend data from quarterly reports: in both cases, manually digitizing chart data is a significant bottleneck that chart-to-table automation would directly eliminate. Despite its relevance, chart-to-table remains a challenging task, and the benchmarks used to evaluate progress on this task share a critical limitation: they do not reflect the diversity of charts found in real scientific publications.

Existing benchmarks such as ChartQA [17], PlotQA [19] and ICPR2022 Chart Competition [3] concentrate on a narrow set of chart types, bar, line, and pie charts, which represent only a fraction of the visual vocabulary used in practice. Statistical charts such as boxplots and error points plots, which appear frequently in biomedical and experimental computer science literature, are systematically absent from these evaluation frameworks. As a result, strong performance on existing benchmarks may present a false picture of generalization: a model that appears capable may have never been exposed to the chart types that scientists actually produce.

A further limitation concerns data authenticity. Most available datasets are either fully synthetic or built around a restricted set of visual templates, producing charts that are significantly cleaner and more stylistically homogeneous than those encountered in real scientific publications. This distribution shift, the mismatch between training and evaluation conditions on one side, and real-world figures on the other, raises serious concerns about whether reported performance translates to practical settings. A model that appears capable on synthetic data may fail when faced with the layout variability and visual noise of actual scientific figures.

To quantify this gap, we conduct a large-scale analysis of chart type distributions across two major scientific repositories: PubMed Central (PMC)[1] and arXiv[2], covering 16,000 papers and over 356,000 extracted chart images. The results confirm that real-world scientific charts are more diverse than current benchmarks reflect.

Motivated by this evidence, we introduce a benchmark and an evaluation metric specifically designed for the chart-to-table task on real scientific figures. Our main contributions are the following: (i) we conduct a large-scale empirical analysis of chart-type distributions across PMC and arXiv, revealing that the visual diversity of real scientific figures substantially exceeds what current benchmarks cover; (ii) we introduce PaperUnPlot, a chart-to-table benchmark with 10 chart categories, including several underrepresented in existing datasets, and 950 manually annotated real chart-table pairs, with 100 samples for each category except bubble charts; (iii) we propose Normalized Mapping Similarity (NMS), an extension of the standard RMS metric for scientific charts with narrow or

Proceedings of the VLDB Endowment. ISSN 2150-8097.

[1] https://pmc.ncbi.nlm.nih.gov/
[2] https://arxiv.org/

logarithmic axes, and use it to systematically assess commercial LVLMs, open-weight multimodal models, and task-specific DePlot. Our evaluation emphasizes that current approaches remain limited when applied to the chart types commonly found in scientific publications.

The benchmark, analysis pipeline, evaluation code, and the full per-chart-type prompts are publicly available [3].

2 RELATED WORK

2.1 Chart Understanding

Chart understanding is a broad research area that encompasses a range of tasks aimed at enabling machines to extract and reason over the information encoded in chart images. The most studied task is **chart question answering** [17] [19] [8] [1] [23] [21] [12] [15] [18] [20], where a model must answer questions in natural language about a chart. Other tasks include **chart summarization** [9] [6], which requires generating a natural language description of the chart content, **chart-to-code** [24] [25], where the goal is to generate the source code that produces a given chart, and **chart segmentation** [16] [5] [3], which focuses on identifying and localizing visual elements within the chart. Despite the diversity of these tasks, the chart-to-table task, which requires reconstructing the underlying data table from a chart image, has received comparatively little attention, and the benchmarks designed for other tasks are rarely suitable for evaluating it directly.

2.2 Chart-to-Table Datasets

Several datasets have been proposed for chart understanding, though few target the chart-to-table task directly, and all share significant limitations in terms of chart type coverage and data authenticity.

ChartQA [17] is a widely used benchmarks for chart question answering, comprising approximately 21,000 real chart images from sources such as Statista and Pew Research, covering bar, line, and pie charts. While its images are real, the chart types and visual conventions do not reflect the distribution of scientific publications, constraining its generalizability to the scientific domain.

PlotQA [19] is a large-scale synthetic dataset of over 224,000 charts, designed for question answering over figures generated from real-world statistical data. Despite its scale, coverage is restricted to bar, line, and scatter charts, and the restricted set of templates raises concerns about generalization to real scientific charts.

ICPR2022 Chart Competition [3] provides approximately 20,000 PubMedCentral chart images annotated on six tasks, with chart-to-table extraction covered by Task 6. The subset annotated for this task covers four chart-type classes (namely: bar, box, line, and scatter charts), making it one of the few existing resources with partial coverage of statistical types. However, chart-type coverage remains limited, and only 5,000 samples carry complete annotations.

ChartMimic [24] is a synthetic benchmark of approximately 4,800 charts primarily designed for the chart-to-code task, covering a wide variety of chart types including histograms, heatmaps, error point plots, and radar charts. Ground truth data tables can in principle be derived from the generating code, though this excludes charts whose values are produced at runtime.

[3] https://github.com/giorgiomelch/about_chart_to_table

ChartBench [23] is a large synthetic chart QA benchmark with 66K charts targeting chart question answering. It covers a range of six chart categories, which are generated synthetically from data collected from Kaggle or generated automatically by GPT.

SimChart9K [22] comprises ~9,300 synthetically generated charts focused on bar and line types, with limited pie chart representation and no statistical chart types. Like other synthetic datasets, it produces figures that are substantially cleaner and more visually uniform than those encountered in scientific literature.

These benchmarks share two fundamental limitations: they focus on a small subset of chart types and a large portion of existing datasets are either fully synthetic or built around restricted visual templates, producing a distribution shift that undermines the validity of evaluations as proxies for real-world performance.

2.3 Evaluation Metrics

Early metrics for chart-to-table evaluation, such as Relative Number Set Similarity (RNSS), treat the predicted output as an unordered set of numeric values, ignoring positional information and non-numeric content. DePlot [11] introduced Relative Mapping Similarity (RMS), which represents a table as a set of header-value triples and computes an optimal matching between predicted and ground truth entries, combining textual similarity on headers with relative numeric distance on values. While RMS represents a significant improvement over RNSS, it presents limitations when applied to scientific charts: it normalizes numeric error by target value magnitude rather than the axis range, and does not account for logarithmic scales. We address both issues with Normalized Mapping Similarity (NMS), introduced in Section 5.

2.4 Chart-to-Table Models

The chart-to-table task has been directly addressed by DePlot [11], a 300M-parameter model that frames it as a sequence-to-sequence problem and fine-tunes a vision-language model to produce linearized tables from chart images. DePlot is trained only on bar, line, and pie charts, which limits its applicability to the broader range of chart types encountered in scientific publications, and serves in this work as a reference point for task-specific approaches. More recently, general-purpose Large Vision-Language Models (LVLMs) have demonstrated competitive zero-shot performance on chart understanding tasks, raising the question of whether task-specific models retain an advantage when evaluated on a broader and more representative benchmark, a question that PaperUnPlot is designed to address.

3 CHART TYPE DISTRIBUTION ACROSS SCIENTIFIC ARTICLES

To quantify the gap between the chart types covered by existing benchmarks and those actually encountered in scientific publications, we conduct a large-scale analysis of figures extracted from two major repositories: PubMed Central (PMC), representative of biomedical literature, and arXiv, representative of database-oriented computer science research.

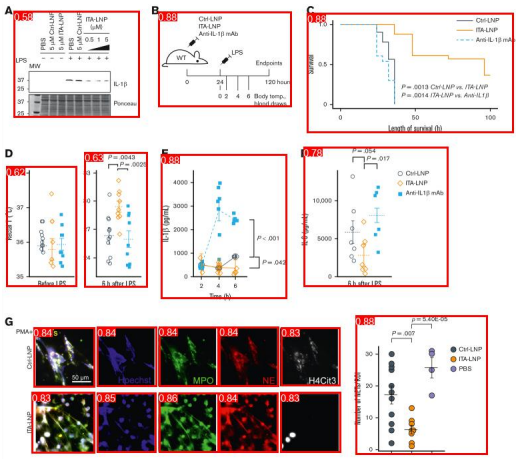

Figure 1: An example of a compound image from PMC.

3.1 Data Collection

We downloaded 8,000 articles from each repository, sampled in chronological order to avoid domain or topic biases introduced by selective queries. From arXiv, articles were retrieved exclusively from the `cs.DB` (Databases) category, ensuring that the sample is representative of database-oriented computer science research. From these articles, we extracted all embedded figures, obtaining **41,256 images from PMC** and **121,585 images from arXiv**. After compound image decomposition (Section 3.2), the total corpus grows to **356,615 individual chart images**.

3.2 Compound Image Decomposition

A distinctive challenge in processing scientific figures, particularly those from PMC, is the widespread presence of *compound images*: figures that contain multiple sub-images arranged in a grid or panel layout (see Figure 1). Since each sub-image may correspond to a different chart type, treating compound figures as atomic units would introduce noise in the classification step. Decomposing them into individual panels is therefore a necessary preprocessing step.

We are not aware of a publicly available dataset for compound image decomposition in scientific figures. We address this in two stages. First, we manually annotated a dataset of **1,000 compound images** with bounding boxes for each sub-image panel, which serves exclusively as our held-out test set. We then evaluated Gemini 3 Flash on this dataset as a baseline in a zero-shot setting, measuring panel-level Precision, Recall, and F1 Score for bounding boxes prediction at Intersection over Unit threshold of 0.5, obtaining a Precision of 0.945, Recall of 0.937, and F1 Score of 0.941.

While Gemini 3 Flash achieves strong zero-shot performance, directly applying it to the full corpus images would be impractical at scale, both in terms of inference cost and reproducibility. Leveraging these results, we used Gemini 3 Flash to automatically annotate **32,000 compound images**, producing a large weakly-supervised training set. We fine-tuned Deformable DETR [27] (two-stage, with iterative bounding box refinement, ResNet-50 backbone) from the official COCO pre-trained checkpoint. Training ran for 10 epochs with AdamW, learning rate 2×10^{-4} (backbone: 2×10^{-5}), weight decay 1×10^{-4}. Batch size is 6, number of object queries is 300.

Evaluated on the manually annotated test set of 1,000 images, the trained model achieves a Precision of 0.927, Recall of 0.930, and F1 Score of 0.928, demonstrating that automatically labeling training data via a large vision-language model is a viable strategy in the absence of curated datasets, while yielding a faster and more cost-effective model for large-scale processing.

We applied this model to decompose all extracted figures from both sources. Processing 41,256 images from PMC yields **204,890 individual sub-images** (an average of 4.97 sub-images per figure), while 121,585 images from arXiv yield **151,725 sub-images** (an average of 1.25 sub-images per figure), resulting in a total corpus of 356,615 individual chart images ready for classification. The substantially higher decomposition ratio observed in PMC highlights the widespread use of multi-panel figures in biomedical literature.

3.3 Chart Classification

We adopt **SwinV2-Base** [14], fine-tuned from an ImageNet-22k [4] checkpoint, as our chart type classifier. The model was trained on a combination of eight publicly available datasets complemented by a manually curated collection, yielding approximately 67,000 labeled images in total, split into training, validation, and test sets following a 70/15/15 ratio. The per-category and per-source breakdown is reported in Table 1. To address class imbalance, underrepresented chart types are augmented with programmatically generated synthetic samples, and a weighted loss is adopted during training. Standard data augmentation is applied, including color jitter, random grayscale conversion, Gaussian blur, and random affine transformations. Training is performed for 12 epochs using the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 64. The classifier achieves an overall accuracy of 98.46% in the test set. To further validate the reliability of the classifier on the target domain, we manually inspected a sample of images drawn from the PMC and arXiv corpora used in the distribution analysis, confirming that classifier predictions were consistent with the true chart types.

3.4 Chart Type Distribution Analysis

We applied the trained classifier to the full corpus of extracted figures from both PMC and arXiv. Table 2 reports the distribution of chart types across PMC and arXiv after classification. A substantial portion of the extracted figures are non-chart images, including diagrams, molecular structures, photographs, and tables, accounting for approximately 52% of PMC figures and 38% of arXiv figures, a difference consistent with the broader use of medical images in biomedical publications. The remaining images, corresponding to scientifically meaningful charts, form the basis of our analysis.

Among chart types, **line** and **bar** charts are by far the most prevalent in both sources, jointly accounting for over 60% of charts in both PMC and arXiv. However, notable differences emerge across domains: PMC exhibits a richer variety of statistical chart types, with box plots (7.94%), error point plots (4.32%), and violin plots (1.94%) appearing at non-negligible frequencies, reflecting the centrality of statistical reporting in biomedical research. In contrast, arXiv figures are more concentrated around line charts (50.2%),

Chart Type	Ours	BlockDiag. [2]	ChartMimic [24]	ChartQA [17]	DocFigure [7]	ICPR22 [3]	MPP [13]	PMC-OA [10]	PubTabNet [26]	Synth.	Total
Area	100	–	154	–	309	–	–	–	–	137	700
Bar	60	–	440	1,000	1,149	10,619	–	–	–	–	13,268
Box	180	–	100	–	566	1,538	–	–	–	–	2,384
Bubble	100	–	–	–	338	–	–	–	–	162	600
Chord Chart	–	–	–	–	–	–	–	–	–	400	400
Contour	210	80	–	360	–	–	–	–	–	150	800
Diagram	–	1,903	–	–	2,097	–	–	–	–	–	4,000
Error Point	190	–	80	–	–	1,257	–	–	–	–	1,527
Heatmap	110	–	120	–	758	172	–	–	–	–	1,160
Histogram	357	–	80	–	–	–	–	–	–	163	600
Image	3,440	–	–	–	4,804	–	–	2,000	–	–	10,244
Line	60	–	280	500	–	14,000	–	–	–	–	14,840
Manhattan	44	–	–	–	–	256	–	–	–	150	450
Map	85	–	–	–	1,052	906	–	–	–	–	2,043
Molecule	50	–	–	–	–	–	2,000	–	–	–	2,050
Pie	85	–	80	541	433	242	–	–	–	100	1,481
Quiver	–	–	57	–	574	–	–	–	–	69	700
Radar/Polar	50	–	80	–	634	–	–	–	–	36	800
Scatter	150	–	84	–	1,132	1,350	–	–	–	–	2,716
Surface 3D	55	–	–	–	392	283	–	–	–	170	900
Table	–	–	–	–	1,898	–	–	–	2,102	–	4,000
Treemap	4	–	80	–	–	–	–	–	–	316	400
Venn	60	–	–	–	862	206	–	–	–	–	1,128
Violin	281	–	80	–	–	–	–	–	–	139	500
Total	**5,671**	**1,983**	**1,715**	**2,401**	**16,998**	**30,829**	**2,000**	**2,000**	**2,102**	**1,992**	**67,691**

Table 1: Number of images per chart category and dataset used to train the SwinV2 classifier.

consistent with the prevalence of performance curves and training plots in computer science publications.

Critically, chart types such as box plots, error point plots, heatmaps, and bubble charts, which appear with meaningful frequency in real scientific publications, are absent or marginal in existing benchmarks such as ChartQA, PlotQA, which focus almost exclusively on bar, line, and pie charts.

Another noteworthy finding is the low frequency of pie charts in both repositories, accounting for only 0.72% of PMC figures and 0.5% of arXiv figures. This stands in contrast with their prominent representation in existing benchmarks, and aligns with a well-established consensus in the data visualization community that pie charts are generally discouraged in scientific publications.

4 PAPERUNPLOT

We introduce PaperUnPlot, a benchmark designed to evaluate chart-to-table models on the diverse range of chart types encountered in real scientific publications. Unlike existing benchmarks, which focus on a small set of common chart types, PaperUnPlot targets categories that are prevalent in scientific literature but are systematically underrepresented in current evaluation frameworks.

4.1 Chart Categories

PaperUnPlot covers 10 chart categories: **bar**, **box**, **bubble**, **error point**, **histogram**, **heatmap**, **line**, **pie**, **radar**, and **scatter**. These categories are selected based on their frequencies in scientific literature, as analyzed in Section 3.4, and because their visual marks can be mapped to discrete tabular entries in a direct and reproducible way. We exclude chart types such as contour plots, quiver plots, 3D surface charts and Manhattan plots, where relevant information is not fully captured by isolated plotted values but also by their arrangement, continuity or domain specific coordinate system. For such charts, multiple tabular annotations may be equally plausible from the same rendered image, making ground-truth definition and metric-based evaluation less reliable.

Chart Type	PMC (%)	arXiv (%)
Line	31.14	50.2
Bar	29.49	29.6
Scatter	9.93	7.0
Box	7.94	3.9
Contour	5.23	0.8
Error Point	4.32	1.1
Heatmap	3.02	2.7
Area	2.74	1.1
Violin	1.94	0.4
Histogram	0.81	1.3
Bubble	1.03	0.0
Venn	0.60	0.3
Radar/Polar	0.38	0.6
Pie	0.72	0.5
Surface 3D	0.50	0.4
Manhattan	0.18	0.0
Quiver	0.04	0.1
Chord Chart	0.0	0.0
Treemap	0.0	0.0

Table 2: Distribution of chart types across PubMed Central and arXiv, sorted by PMC frequency. Percentages refer to individual chart images after compound image decomposition and classification.

4.2 PaperUnPlot Data Collection

For each selected category, we collected 100 chart images sourced from the same pool of 8,000 papers per repository used in the distribution analysis. Images are drawn equally from PubMed Central (PMC) and arXiv, 50 charts per category from each source, to ensure broad coverage across scientific domains. The only exception is bubble charts: as confirmed by the distribution analysis in Table 2, bubble charts are virtually absent from arXiv (0.0%), resulting in 50 samples collected exclusively from PMC for this category. The total dataset comprises **950 real chart images**.

Charts were sampled to reflect the visual diversity encountered in practice, including varying styles, color schemes, fonts, axis

configurations, and levels of visual complexity. No filtering was applied to exclude charts that are difficult to read or annotate, in order to preserve the realistic difficulty of the task.

4.3 Annotation

Each chart image is paired with a manually produced tabular annotation that represents the underlying data. Annotations were created with the assistance of `WebPlotDigitizer`,[4], a tool that enables precise extraction of data points from chart images through interactive calibration of the axes system. For each chart, annotators identified axis ranges and scale, extracted all data points, and organized them into a structured JSON file. Each annotation explicitly records the scale type of each axis, distinguishing between linear and logarithmic scales, a field that is directly exploited by the NMS metric to apply the appropriate distance function during evaluation. For chart types that encode values through color, such as heatmaps, we developed a dedicated script that maps pixel colors to data values by calibrating each chart's colorbar. All annotations were reviewed for correctness.

4.4 Synthetic Split

In addition to the real split, we generated a synthetic counterpart for each chart in the benchmark by rendering the corresponding ground truth table as a chart using matplotlib, varying color scheme and layout across samples. This yields a synthetic split of 950 charts in a one-to-one correspondence with the real split, each synthetic chart encodes exactly the same underlying data as its real counterpart. This paired design enables a controlled comparison of model performance on real versus synthetic charts, isolating the effect of visual complexity from data complexity: any performance gap between the two splits can be attributed directly to the visual diversity of real scientific figures rather than to differences in the underlying data.

4.5 Benchmark Statistics

Table 3 reports the composition of the real split by chart type and source. Figure 2 shows the distribution of data points per chart type across the benchmark. Histogram, scatter and heatmap charts exhibit the highest median number of data points, reflecting the density of measurements typically reported in these formats, while pie, bar and box tend to encode fewer elements.

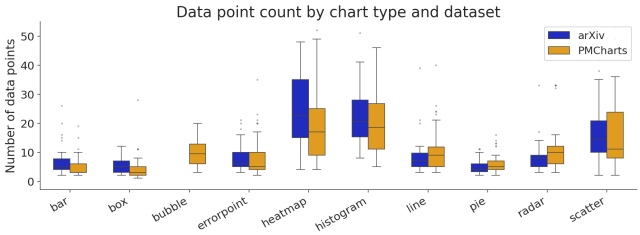

Figure 2: Distribution of data points per chart type across PaperUnPlot.

[4]https://github.com/automeris-io/WebPlotDigitizer

Chart Type	PMC	arXiv	Total
Bar	50	50	100
Box	50	50	100
Bubble	50	0	50
Error Point	50	50	100
Histogram	50	50	100
Heatmap	50	50	100
Line	50	50	100
Pie	50	50	100
Radar	50	50	100
Scatter	50	50	100
Total	**500**	**450**	**950**

Table 3: Composition of PaperUnPlot by chart type and source repository.

5 EVALUATION METRIC

Evaluating the quality of a predicted table against a ground truth table is a non-trivial problem. Early metrics such as Relative Number Set Similarity (RNSS) consider only unordered set of numeric entries, ignoring the position of values within the table and all non-numeric content. To address this, DePlot introduced Relative Mapping Similarity (RMS), which we adopt as our baseline metric.

RMS represents a table not as a flat set of numbers but as an unordered collection of mappings from row and column headers (r, c) to a value v. Each entry is a triple $p_i = (p_i^r, p_i^c, p_i^v)$ for a predicted table $P = \{p_i\}_{1 \le i \le N}$ and $t_j = (t_j^r, t_j^c, t_j^v)$ for a target table $T = \{t_j\}_{1 \le j \le M}$. The distance between two entries combines a textual component, measured via Normalized Levenshtein Distance NL_τ on the concatenated headers, and a numeric component:

$$D_{\tau,\theta}(p, t) = \left(1 - NL_\tau(p^r \| p^c,\ t^r \| t^c)\right)\left(1 - D_\theta(p^v, t^v)\right) \quad (1)$$

where $D_\theta(p, t) = \min(1, |p - t|/|t|)$ is the relative distance between numeric values, and thresholds τ and θ prevent partial credit for very dissimilar entries. Following the original DePlot evaluation protocol, we set $\tau = 0.5$ and $\theta = 0.1$ throughout all experiments. An optimal matching $X \in \mathbb{R}^{N \times M}$ between entries is computed based on header similarity, and precision and recall are defined as:

$$\mathrm{RMS}_{\mathrm{precision}} = 1 - \frac{\sum_i \sum_j X_{ij}\, D_{\tau,\theta}(p_i, t_j)}{N} \quad (2)$$

$$\mathrm{RMS}_{\mathrm{recall}} = 1 - \frac{\sum_i \sum_j X_{ij}\, D_{\tau,\theta}(p_i, t_j)}{M} \quad (3)$$

The final $\mathrm{RMS}_{\mathrm{F1}}$ is the harmonic mean of precision and recall. The metric is invariant to row and column permutations, and transpositions are handled by returning the score corresponding to the best-matching orientation.

Limitations and proposed modifications. The relative distance $D_\theta(p, t) = \min(1, |p - t|/|t|)$ used in the original RMS presents two limitations when applied to scientific charts. First, it normalizes the error by the magnitude of the target value $|t|$ rather than by the actual range of the axis. This can lead to a failure case: consider a chart whose axis spans values from 1000 to 1001, a range of just

one unit. If the model predicts 1000 instead of 1001, the relative distance is $\min(1, 1/1001) \approx 0.001$, suggesting near-perfect accuracy. Yet in the context of that chart, the model has missed the correct value by the entire extent of the visible axis, a substantial error that the metric fails to capture. Normalizing by the axis range $axis_{\max} - axis_{\min}$ instead correctly assigns this prediction a maximum distance of 1. The second limitation concerns logarithmic axes, which are common in scientific publications when data spans several orders of magnitude. On a logarithmic scale, equal visual intervals correspond to multiplicative rather than additive differences: the distance between 1 and 10 is visually equivalent to the distance between 10 and 100. Applying a linear distance measure on such axes misrepresents the actual magnitude of prediction errors relative to what a reader would perceive from the chart.

We address both issues by replacing the relative distance with two axis-aware variants. For linear axes, we adopt a normalized distance bounded by the axis range:

$$D_\theta(p, t) = \min\left(1, \frac{|p - t|}{axis_{\max} - axis_{\min}}\right) \tag{4}$$

For logarithmic axes, we instead use a logarithmic distance that operates in log-space and normalizes by the log-extent of the axis:

$$D_\theta(p, t) = \min\left(1, \frac{|\log p - \log t|}{|\log(axis_{\max}) - \log(axis_{\min})|}\right) \tag{5}$$

All other components of RMS remain unchanged. We refer to this extended metric as **Normalized Mapping Similarity (NMS)**. Applying NMS to existing benchmarks is not straightforward, as our method relies on annotations with explicit axis ranges and scale types (linear or logarithmic), which are not provided in those datasets.

6 EXPERIMENTS

We evaluate eight models spanning three categories.

Commercial models. The first category comprises four commercial LVLMs: **GPT-4o**, **GPT-4o Mini**, **Gemini 2.5 Flash** and **Claude Sonnet 4.5**.

Open-weight models. The second category comprises three open-weight multimodal models in the 2–4B parameter range: **Qwen2-VL 2B**, **Phi-3.5 Vision** (4.2B), and **InternVL2 2B**. These models are selected to assess whether small open-weight models, deployable on commodity hardware without commercial API access, represent viable alternatives for chart-to-table extraction in resource-constrained scenarios.

Task-specific model. The third category comprises a single task-specific model, **DePlot** [11], a 300M-parameter model described in Section 2.

6.1 Experimental Setup

All models except DePlot are prompted with a chart-type-specific prompt designed to capture the structural differences between chart categories in the expected output format. Each image is resized to 768 pixels on its longest side. Prompts share a common base structure instructing the model to act as an expert data analyst, extract the most accurate values possible based on the visible axes, and estimate values when exact labels are not present. Output is required in valid JSON.

The base output structure includes `chart_title`, `x_axis_label`, `y_axis_label`, and a list of `data_points`, each with `series_name`, `x_value`, and `y_value`. For chart types encoding statistical distributions, the `y_value` field is expanded: box plots use a nested object with `min`, `q1`, `median`, `q3`, and `max`; analogous adaptations apply to bubble, error point, and heatmap charts.

When model outputs cannot be parsed as valid JSON, a failure mode observed only among open-weight models, the prediction is treated as an empty table, contributing zero to both precision and recall.

6.2 Results

Figure 3 reports NMS F1 scores for all eight models across the two real splits and the synthetic split of PaperUnPlot. Three patterns emerge consistently: commercial models substantially outperform open-weight models; Gemini 2.5 Flash and Claude Sonnet 4.5 are the best-performing models overall; and performance on synthetic charts is uniformly higher than on real charts.

Commercial models. Among commercial models, Gemini 2.5 Flash achieves the best overall performance across both real splits, followed by Claude Sonnet 4.5, with a large and consistent gap separating this top two from GPT-4o and GPT-4o Mini. This ranking holds across both arXiv and PMC. The difference between the best and worst commercial models is substantial, with Gemini outscoring GPT-4o Mini by more than 40 points on both arXiv (+44.2) and PMC (+40.9).

A different ordering emerges on the synthetic split, where Claude overtakes Gemini. This shift in ranking is driven by Claude's large performance gain when moving from real to synthetic charts (+9.7 on arXiv, +14.7 on PMC), contrasted with Gemini's near-identical scores across the two settings (+0.7 and +0.1). This suggests that Gemini is substantially more robust to the visual variability of real scientific figures, while Claude's stronger performance on clean, template-based charts does not fully transfer to the more diverse layouts encountered in actual publications.

Open-weight models. Open-weight models fall substantially behind commercial models across both real splits. The best performer in this group, Phi-3.5 Vision, approaches GPT-4o Mini on both splits despite being considerably smaller, while InternVL2 2B and Qwen2-VL 2B score well below it. Notably, all three open-weight models achieve relatively stronger results on pie charts and heatmaps, the two categories where explicit value labels reduce the task largely to OCR, yet remain far from competitive with commercial models even on these simpler types.

A qualitative analysis of model outputs reveals several recurring failure modes. All three models tend to undercount the elements present in a chart: Phi-3.5 Vision typically reports only one series in multi-series charts, silently omitting the others, while Qwen2-VL 2B predicts only a fraction of the data points. Structural confusions are also common: Qwen2-VL 2B conflates data series with axis labels and axis labels with data values, whereas InternVL2 2B frequently returns well-formed JSON with correctly identified elements but

null values, indicating that structural parsing succeeds where numerical reading fails. InternVL2 2B also struggles to associate values to elements that fall between axis ticks, falling back to the nearest tick or midpoint, and is sensitive to visual clutter, when bar or error point charts contain overlaid scatter points, it extracts the scatter points rather than the intended elements. Finally, on structurally complex chart types such as box plots and bubble charts, where each element requires extracting multiple values, Qwen2-VL 2B and InternVL2 2B tend to reduce each element to a single value.

Beyond qualitative errors, a non-negligible fraction of open-weight model outputs are absent or unparseable: InternVL2 2B on 9.2% of charts, Phi-3.5 Vision on 4.5%, and Qwen2-VL 2B on 3.5%. These failures took two distinct forms: either the model ignored the required JSON structure, returning an invalid schema, or it produced a truncated output, continuously appending extracted elements until the maximum token limit was reached. Failures were not uniformly distributed: InternVL2 2B alone failed on 58% of heatmap instances, and across all three models, unparseable outputs were concentrated on charts with many elements to extract, a median of 20 data points, versus 12 for successfully parsed ones.

These patterns suggest that open-weight models at this scale lack the visual grounding necessary for reliable chart-to-table extraction, particularly on charts with dense or structurally complex data.

Task-specific model. DePlot performs poorly across both real splits (arXiv: 21.2, PMC: 14.7), with six chart categories contributing a score of zero, as DePlot was trained exclusively on bar, line, and pie charts. Histogram is a partial exception: its visual similarity to bar charts allows DePlot to extract some structure despite never having been explicitly trained on it. Even on the three categories DePlot was trained on, it fails to achieve competitive scores to commercial models: on the arXiv split, bar (62.5 vs 93.5 for Gemini), line (55.3 vs 91.2), and pie (50.2 vs 94.5). The gap widens further on PMC. This suggests that its limitations extend beyond chart type coverage and reflect a broader inability to generalize to the visual diversity of real scientific figures. This interpretation is supported by DePlot's performance on the synthetic split, where scores are substantially higher across all trained categories: bar (+19.7 arXiv, +26.8 PMC), line (+19.9 arXiv, +34.7 PMC), and pie (+25.0 arXiv, +16.7 PMC). The gap between synthetic and real performance is larger for DePlot than for any other evaluated model, providing direct evidence that it has overfit to a clean, template-based visual style of its training data and fails to transfer to the more varied figures encountered in actual publications.

A qualitative inspection of DePlot outputs reveals several recurring failure modes. The model is sensitive to visual noise: labels or annotations placed near chart elements are frequently misinterpreted as data values, for instance, a text annotation adjacent to a bar may be read as the bar's value. For pie charts where the legend associates colors with slices rather than labeling them directly, the model frequently fails to correctly associate slice labels with their values, a layout likely underrepresented in training. On charts with two continuous numeric axes, such as histograms, the model struggles to extract values in the absence of categorical labels.

Cross-category analysis. Radar and bubble charts represent the hardest category across all models and both real splits. The primary difficulty for radar charts lies in incorrect association of values to categories: since data is encoded along angular axes radiating from a central point, models systematically struggle to map values to the correct spoke, particularly for categories located far from a reference axis. Bubble charts present a distinct difficulty: beyond identifying the x and y position of each bubble, models must additionally estimate the bubble's size and its color, two visual channels that prove harder to decode than positional information.

Histograms also prove challenging, particularly when the number of bins is large. Models tend to undercount bins, merging adjacent bars into a single entry or omitting bars near the axis extremes. The lack of explicit bar labels adds to this difficulty, leaving models to estimate values from bar height alone.

In contrast, heatmaps and pie charts rank among the easier categories across most models. Heatmaps typically display numeric values directly inside each cell, reducing a large part of the task to OCR. Similarly, pie charts frequently annotate each slice with its percentage value, limiting difficulty to correctly associating each label with the right slice.

Synthetic versus real-world performance. Across all models, performance on the synthetic split is higher than on the corresponding real categories, confirming a distribution shift between clean, template-based figures typically used in training and evaluation and visually diverse charts encountered in actual publications.

	Input		Output		
Model	Avg Tok.	$/1M	Avg Tok.	$/1M	Total Cost ($)
GPT-4o	1212	$2.50	498	$10.00	$15.21
GPT-4o-mini	1212	$0.15	398	$0.60	$0.80
Gemini 2.5 Flash	500	$3.50	605	$2.50	$6.20
Claude Sonnet 4.5	1040	$3.00	531	$15.00	$21.05

Table 4: Comparative analysis of inference costs and average token consumption across commercial models (N=1900 charts).

6.3 Cost Analysis

Table 4 reports the estimated inference cost for each commercial model evaluated on the full PaperUnPlot benchmark (950 real charts and 950 synthetic charts, for a total of 1900 chart images). Costs are computed based on official pricing at the time of evaluation,[5] using the average token counts observed across our experiments. Gemini 2.5 Flash achieves the best performance at the second lowest cost among commercial models, representing the most favorable performance-to-cost trade-off. Claude Sonnet 4.5, while the second best performer, incurs the highest cost due to its elevated output token pricing. GPT-4o Mini offers the lowest absolute cost but at a substantial performance penalty relative to the other commercial models.

7 CONCLUSION

We presented PaperUnPlot, a dataset of 950 real chart-table pairs spanning 10 chart categories drawn from PubMed Central and arXiv. Alongside it, we conducted a large-scale analysis of chart type distributions across two major scientific repositories, revealing a substantial gap between the charts prevalent in real publications

[5]Prices retrieved on 06/05/2026.

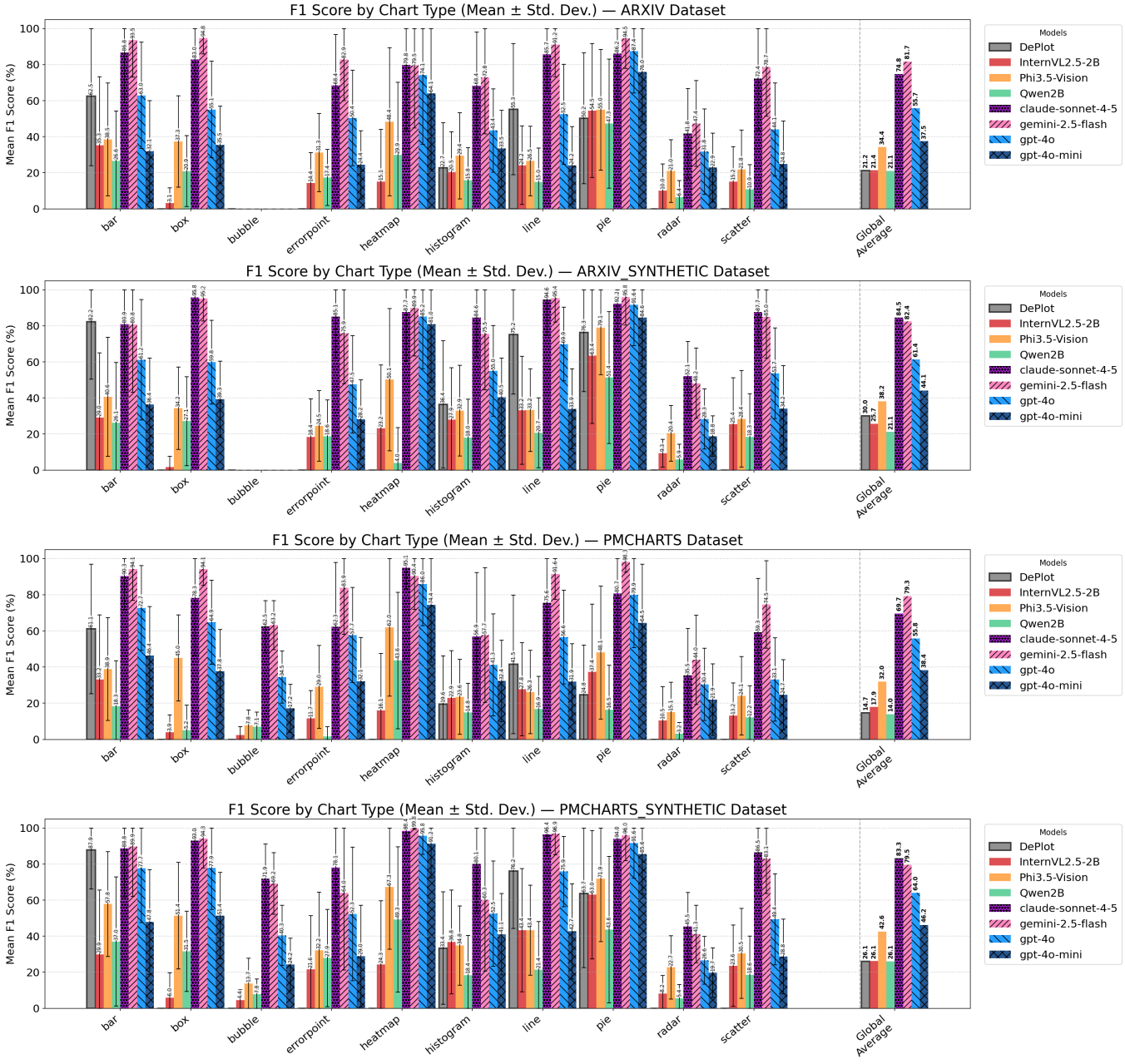

Figure 3: NMS F1 score for each model across the 10 chart categories of PaperUnPlot and the synthetic split.

and those covered by existing datasets. We evaluated eight models on PaperUnPlot, covering commercial LVLMs, open-weight multimodal models, and the task-specific model DePlot. Our results show that performance varies substantially across chart types, that open-weight models fall well short of commercial ones, and that synthetic datasets systematically overestimate practical model capabilities, a finding consistent across all evaluated models.

Future work includes increasing the number of samples per category, fine-tuning task-specific and open-weight models on visually diverse real-world figures, and exploring alternative prompting strategies such as chain-of-thought reasoning or multi-step agentic pipelines, where a model first identifies the chart structure before proceeding to data extraction. Systematically characterizing task difficulty along dimensions such as chart type, visual template, number of data series, and element count would also provide a principled basis for analyzing model limitations and guiding future benchmark design.

REFERENCES

[1] Saleem Ahmed, Bhavin Jawade, Shubham Pandey, Srirangaraj Setlur, and Venu Govindaraju. 2023. Realcqa: Scientific chart question answering as a test-bed for first-order logic. In *International Conference on Document Analysis and Recognition*. Springer, 66–83.

[2] Shreyanshu Bhushan and Minho Lee. 2022. Block diagram-to-text: Understanding block diagram images by generating natural language descriptors. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*. 153–168.

[3] Kenny Davila, Fei Xu, Saleem Ahmed, David A Mendoza, Srirangaraj Setlur, and Venu Govindaraju. 2022. ICPR 2022: challenge on harvesting raw tables from infographics (chart-infographics). In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 4995–5001.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[5] Muhammad Yusuf Hassan, Mayank Singh, et al. 2023. LineEX: Data extraction from scientific line charts. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 6213–6221.

[6] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 3258–3264.

[7] KV Jobin, Ajoy Mondal, and CV Jawahar. 2019. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 1. IEEE, 74–79.

[8] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. Opencqa: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 11817–11837.

[9] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4005–4023.

[10] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 525–536.

[11] Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 10381–10399.

[12] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1287–1310.

[13] Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. 2025. FGBench: A Dataset and Benchmark for Molecular Property Reasoning at Functional Group-Level in Large Language Models. *arXiv preprint arXiv:2508.01055* (2025).

[14] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12009–12019.

[15] Yujing Lu, Ling Zhong, Jing Yang, Weiming Li, Peng Wei, Yongheng Wang, Manni Duan, and Qing Zhang. 2026. DomainCQA: Crafting Knowledge-Intensive QA from Domain-Specific Charts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 32347–32355.

[16] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1917–1925.

[17] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*. 2263–2279.

[18] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*. 19123–19151.

[19] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*. 1527–1536.

[20] Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Sprague, et al. 2025. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *arXiv preprint arXiv:2505.13444* (2025).

[21] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems* 37 (2024), 113569–113697.

[22] Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, and Junchi Yan. 2023. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268* (2023).

[23] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915* (2023).

[24] Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. 2024. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961* (2024).

[25] Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. Boosting chart-to-code generation in mllm via dual preference-guided refinement. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 11032–11041.

[26] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*. Springer, 564–580.

[27] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).