SciVQA 2025: Overview of the First Scientific Visual Question Answering Shared Task

Ekaterina Borisova[1,2], **Nikolas Rauscher**[1,2], **Georg Rehm**[1,3]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
[2]Technische Universität Berlin [3]Humboldt-Universität zu Berlin
Corresponding author: ekaterina.borisova@dfki.de

Abstract

This paper provides an overview of the First Scientific Visual Question Answering (SciVQA) shared task conducted as part of the Fifth Scholarly Document Processing workshop (SDP 2025). SciVQA aims to explore the capabilities of current multimodal large language models (MLLMs) in reasoning over figures from scholarly publications for question answering (QA). The main focus of the challenge is on closed-ended visual and non-visual QA pairs. We developed the novel SciVQA benchmark comprising 3,000 images of figures and a total of 21,000 QA pairs. The shared task received seven submissions, with the best performing system achieving an average F1 score of approx. 0.86 across ROUGE-1, ROUGE-L, and BertScore metrics. Participating teams explored various fine-tuning and prompting strategies, as well as augmenting the SciVQA dataset with out-of-domain data and incorporating relevant context from source publications. The findings indicate that while MLLMs demonstrate strong performance on SciVQA, they face challenges in visual reasoning and still fall behind human judgments.

1 Introduction

Graphical representations such as figures (e. g., charts and diagrams), combined with natural language, serve as essential tools for identifying patterns, analysing trends, and extracting insights from data. In academic research, this dual-modality is particularly prominent, with scientific publications conveying large amounts of valuable information through both unstructured text and (semi-)structured figures.

Automatically decoding and processing data from figures available in scholarly papers (i. e., *scientific figures*) can be beneficial for downstream tasks such as visual question answering (VQA). However, VQA over figures is challenging due to their diverse types (e. g., line charts, box plots, pie charts), multimodal nature (combining visuals, numerical data, text), and complex relationships between various components (e. g., axes and labels) (Meng et al., 2024; Zhou et al., 2023). For scientific figures, the task is further complicated by the presence of domain-specific terminology and principles (Huang et al., 2024). Hence, efficient VQA requires accurate information extraction, strong reasoning skills, and expertise in the target research field (Liu et al., 2023b; Li et al., 2024b; Meng et al., 2024).

Although VQA has been extensively studied (Wu et al., 2017), its application to scientific figures is still an emerging area of research (Ahmed et al., 2023). Existing real-world datasets are limited, containing figures sourced exclusively from arXiv [1] (Wang et al., 2024b; Roberts et al., 2024), ignoring other scientific contexts, such as peer-reviewed conference and journal publications. Furthermore, while several works examine the robustness of current multimodal large language models (MLLMs) for figure VQA (Islam et al., 2024; Mukhopadhyay et al., 2024; Wu et al., 2024), none specifically focus on extensive evaluation of models' abilities to accurately recognise, process, and link visual attributes (e. g., colour, shape, size) of scientific figures with textual content (e. g., captions, legends, axis labels).

To bridge the mentioned gaps and promote further research, we organised the *First Scientific Visual Question Answering* (SciVQA) shared task as part of the Fifth Scholarly Document Processing workshop (SDP 2025)[2] at ACL 2025. This challenge aims to shed light on the capabilities and limitations of current MLLMs in handling both

[1] https://arxiv.org

[2] https://sdproc.org/2025/

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 182–210
July 31, 2025 ©2025 Association for Computational Linguistics

questions addressing visual elements of scientific figures and those without visual information. Participants were invited to build VQA systems using a novel dataset of 3,000 images of scientific figures from two distinct sources, ACL Anthology[3] and arXiv, associated with a total of 21,000 visual and non-visual QA pairs. The competition attracted 20 registered teams, seven of which submitted their results. This paper presents an overview of the SciVQA shared task, including the dataset, baseline, and submitted systems description, summary of the results, comparison of automatic solutions to human performance, and an analysis of common challenges and errors faced by MLLMs.

2 Related work

Existing datasets. Previous efforts such as FigureQA (Kahou et al., 2018), DVQA (Kafle et al., 2018), LEAF-QA (Chaudhry et al., 2019), and PlotQA (Methani et al., 2020), rely on synthetic data with limited types of figures and template-based QA pairs. For instance, FigureQA focuses on bar, line, and pie charts plotted using the Bokeh library and associated with QA pairs generated from the fifteen predefined templates. DVQA is even more restricted in terms of figure variability, containing only bar plots generated with the Matplotlib library. While such datasets utilise the low-cost approach for data generation and annotation, they fail to reflect complexity and diversity of real-world figures and questions. Current benchmarks, including ChartQA (Masry et al., 2022), OpenCQA (Kantharaj et al., 2022), CharXiv (Wang et al., 2024b), SciFiBench (Roberts et al., 2024), and ChartQAPro (Masry et al., 2025a), comprise authentic images of figures with either human-written or manually validated synthetic QA pairs. However, only the latter three feature unbounded types of figures. Additionally, existing datasets vary in terms of the QA taxonomies they adopt. Among the commonly distinguished question categories are structural (understanding a figure's structure), retrieval (extracting information from a figure's components), and reasoning (operating on multiple figures' components), with binary (yes/no), multiple-choice, fixed or open vocabulary answers (Kafle et al., 2018; Chaudhry et al., 2019; Methani et al., 2020; Masry et al., 2022, 2025a). Recent works, CharXiv and ChartQAPro, also introduce the novel distinction between answerable and unanswerable questions. Although diverse benchmarks are available, those containing real-world scientific figures and questions remain scarce and are primarily limited to a single source – pre-prints from arXiv.

Modeling approaches. Earlier studies (Liu et al., 2023a; Kim et al., 2020; Masry et al., 2022; Methani et al., 2020; Liu et al., 2023b; Zhou et al., 2023) approach QA over figures with a two-stage process, i. e., the image of a figure is transformed into an underlying (semi-)structured table which then serves as part of a textual input to a language model. One of the main drawbacks of this method is the loss of visual information such as colour (e. g., purple box), shape (e. g., triangular marker), position (e. g., top right figure), height (e. g., between the highest and the lowest bars), direction (e. g., pointing toward the box), and size (e. g., largest segment) (Liu et al., 2023a; Kim et al., 2024; Wei et al., 2024), which prevents systems from answering questions that rely on these features (e. g., *"What is the minimum value of the **green line**?"*). With recent advances in vision and multimodality research, the focus has shifted towards an end-to-end VQA approach, i. e., leveraging images of figures directly using MLLMs, thus preserving visual aspects (Wang et al., 2024b; Masry et al., 2025b; Han et al., 2023; Zeng et al., 2024; Wei et al., 2024). While some works propose and utilise figure-oriented MLLMs, including ChartGemma (Masry et al., 2025b), ChartLlama (Han et al., 2023), UniChart (Masry et al., 2023), ChartAssistant (Meng et al., 2024), TinyChart (Zhang et al., 2024), and MultiModal Chart Assistant (Liu et al., 2024), others (Mukhopadhyay et al., 2024; Wu et al., 2024) also explore the capabilities of general-purpose MLLMs such as GPT-4o (OpenAI et al., 2024) and Gemini (Team et al., 2024) via prompt engineering. Despite the promising results of the current open- and closed-source MLLMs in VQA over figures (Islam et al., 2024; Mukhopadhyay et al., 2024; Wu et al., 2024), their effectiveness in accurately recognising and interpreting visual attributes (e. g., colour, shape, height) remains underexplored.

Compared to the existing works, the SciVQA shared task is intended to advance VQA over scientific figures, specifically focusing on exploring the capabilities of MLLMs to reason over questions addressing visual aspects of objects such as shape, size, position, height, direction or colour.

[3] https://aclanthology.org

3 Shared task overview

In the SciVQA challenge, the task is to develop multimodal QA systems using images of scientific figures, their captions, associated natural language QA pairs, and optionally additional metadata (e. g., figure type). The shared task was hosted on the Codabench platform (Xu et al., 2022) from April 1, 2025, to May 16, 2025.[4] In what follows, QA pair types schema (§3.1), dataset (§3.2), and metrics used for evaluation (§3.3) are described in detail.

3.1 Question answering pair types schema

As mentioned in §2, prior studies mainly rely on fixed templates for QA pairs generation. However, this approach restricts the diversity and naturalness of the resulting QA pairs. Due to these limitations, we defined a custom schema containing seven QA pair types. As shown in Figure 1, the QA pairs fall into two root classes: *closed-ended* and *unanswerable*. A closed-ended QA means that it is possible to answer a question based solely on a given data source, i. e., a figure image and/or optionally its caption. Thus, no additional resources such as the main text of a publication, other documents, figures or tables are required. In contrast, an unanswerable question implies that it is not possible to infer an answer solely from a given data (e. g., full paper text is required, values are not visible or missing).

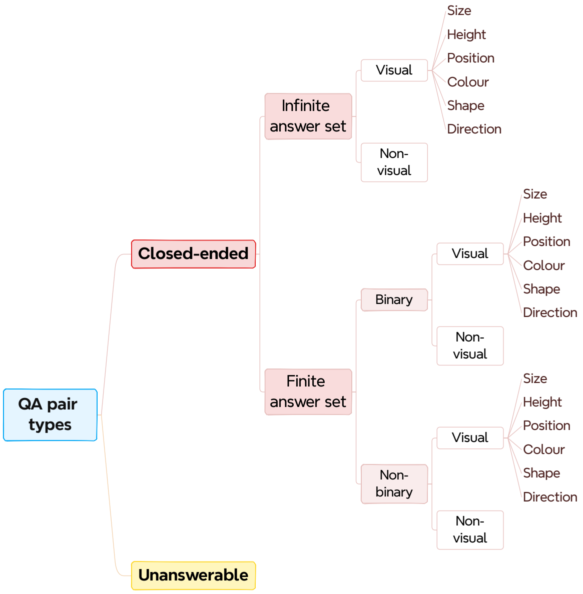


Figure 1: Question answering pair types schema.

At the second level of our schema, the categorisation is based on the fact that for a given question Q, there exists a set S of all possible answers $S = \{a_1, a_2, \ldots, a_N\}$, which can be either *infinite* or *finite*. Questions with an infinite S of answers simply do not have any predefined answer options, e. g., *"What is the approximate value of the loss at the 10th epoch for the green line?"*. On the contrary, questions with a finite S of answers are associated with a limited range of answer options. Such QA pairs fall into two subcategories: 1. *binary* – require a yes/no or true/false answer, e. g., *"Is the percentage of positive tweets equal to 15?"*; 2. *non-binary* – require to choose from a set of M predefined answer options where one or more are correct, e. g, *"What is the maximum value of the green bar at the threshold equal to 10?" – A: 5, B: 10, C: 300, D: None of the above*. Each of the discussed QA pair types can be *visual* and *non-visual*. Visual questions address or incorporate information on one or more of the six visual attributes of a figure, i. e., shape, size, position, height, direction or colour, e. g., *"In the **bottom left** figure, what is the value of the **blue line** at iteration 100?"*. Non-visual questions do not involve any of the mentioned six visual aspects of a figure, e. g, *"What is the minimum value of X?"*, *"What is the difference between the percentage of votes obtained for humour and non-humour tweets?"*. Table 3 (Appendix A) summarises QA pair types and their definitions, while Figure 4 (Appendix A) provides an example of an annotated figure.

3.2 SciVQA dataset

Data collection. The SciVQA dataset comprises 3,000 images[5] of real-world figures extracted from English scientific publications in Computational Linguistics (CL). The figure instances are collected from the two existing datasets, ACL-Fig (Karishma et al., 2023) and SciGraphQA (Li and Tajbakhsh, 2023). ACL-Fig is a corpus of 1,671 figure images extracted from ACL Anthology papers and automatically annotated for the type classification task. SciGraphQA is a dataset of 295,000 figure images from scholarly publications available on arXiv, annotated for multi-turn VQA. First, we extract all figures from the ACL-Fig dataset, excluding images of tables and those not depicting any trends or consisting solely of text, i. e., instances classified as algorithms, natural images, NLP rules/grammar, screenshots, maps, and word clouds. Then to obtain the remaining data, we take a random sample of

[4] https://www.codabench.org/competitions/5904/

[5] We restrict the dataset size due to constraints in both the annotation timeframe and the number of annotators available.

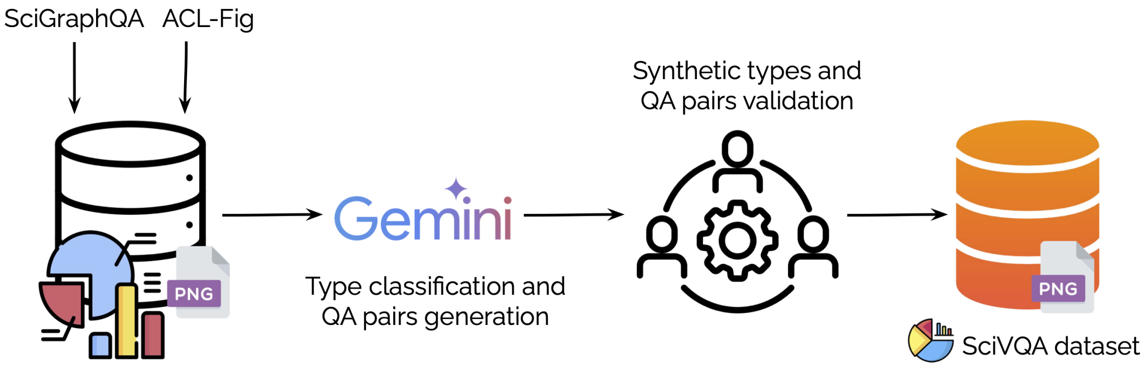

SciGraphQA
ACL-Fig
PNG
Gemini
Type classification and QA pairs generation
Synthetic types and QA pairs validation
PNG
SciVQA dataset

Figure 2: SciVQA dataset annotation pipeline.

figures from SciGraphQA that originated from papers tagged as Computation and Language (cs.CL). We also perform deduplication to ensure that only figures from papers not already included in the ACL-Fig subset are considered. Finally, we manually assess the quality of all collected images and substitute those which are unreadable due to low resolution or issues such as fully cropped y and x labels. We use PDFFigures 2.0 (Clark and Divvala, 2016) and MinerU (Wang et al., 2024a) to extract the respective figure images from the PDF files of scholarly papers. As a result, the SciVQA corpus contains 908 images from ACL-Fig and 2,092 images from SciGraphQA, fetched from scholarly papers published between 1994 and 2024.

Annotation. Inspired by recent studies (Li and Tajbakhsh, 2023; Li et al., 2024a) which leverage generative models like GPT-4 for generating QA pairs, we annotate the SciVQA dataset semi-automatically to reduce the manual effort and cost. The annotation process involves two main phases (see Figure 2): 1. synthetic QA pairs generation and figure type classification 2. followed by manual validation of the results.

In the initial stage, we perform automatic annotation based on figure images and captions using the free API tier of Gemini-1.5-Flash (Team et al., 2024).[6] First, we classify figures into types according to structural and stylistic characteristics, as this information can serve as useful metadata during VQA systems training. All figures are classified as either *compound*, i. e., contain multiple sub-figures which can be separated and constitute individual figure objects or *non-compound*, i. e., contain a single figure which cannot be decomposed into multiple standalone sub-figures. Additionally, we instruct Gemini to indicate the number of (sub-)figures in a given image. Following the ACL-Fig schema, we further categorise the figures in the SciGraphQA subset into one of the following eleven types: *line chart*, *bar chart*, *box plot*, *confusion matrix*, *pie chart*, *scatter plot*, *pareto chart*, *venn diagram*, *architecture diagram*, *neural networks*, and *tree*. Note that we exclude the *graph* class, as it is too generic and the model might overuse it. However, we retain it during the human validation phase since figures from ACL-Fig already include this label. Finally, we annotate each figure image from SciVQA with seven synthetic QA pairs according to the schema discussed in §3.1. For the unanswerable questions, we instruct the model to output the predefined statement *"It is not possible to answer this question based only on the provided data."*, and to generate four answer options for non-binary questions.[7] As a result, a total of 21,000 QA pairs are obtained. Prompt examples are provided in Figures 5-7 in Appendix B.

In the next phase, we manually validate synthetic QA pairs and figure type labels. We hire five master students with a strong theoretical background in CL and a high level of proficiency in English. We also involve three additional student assistants from our lab with the relevant expertise. As an annotation tool, we use Label Studio[8] since it allows both image and text input. Depending on their contracts, each annotator is assigned 133-520 images, i. e., 931-3,640 QA pairs. To mitigate potential bias from an annotator working primarily on a single figure type (e. g., line graph), we ensure that each student receives a diverse set of figures. In the annotation setup, students are pro-

[6] The use of Gemini-1.5-Flash was prohibited during the competition to eliminate any bias.

[7] The data preparation code is available in our GitHub repository: https://github.com/esborisova/SciVQA

[8] https://labelstud.io

vided with a figure image, its caption, type labels, and seven QA pairs with information on their types (see Figure 11 in Appendix C). For the figure classification, we also introduce an *other* category to account for instances outside the ACL-Fig schema. Annotators could specify a subclass if they know the specific type. Additionally, there is an option to request access to the source PDF file. However, since the task requires questions to be answerable without additional context, the annotators are instructed to consult the corresponding PDF file only in edge cases (e. g., unclear or unfamiliar terminology). The students are asked to either confirm or edit the synthetic annotations based on the evaluation criteria defined in the guidelines.[9] Note that no inter-annotator agreement is computed, as each data instance is validated by one student. The annotation project lasted for two months, including one week of training during which the annotators familiarised themselves with the guidelines, Label Studio, and completed a trial annotation of 20 images. As a result, 14,013 out of 21,000 (i. e., about 67%) QA pairs generated by Gemini are modified during this phase.

Each data point in the final annotated SciVQA dataset includes the PNG file of a figure and metadata such as QA pair, QA pair type, caption text, instance ID, image filename, figure ID, figure type labels, number of (sub-)figures, source paper ID and URL, venue, field (for arXiv data), and source dataset. The resulting corpus is split into train (70%), validation (10%), and test (20%) sets (see Table 4 in Appendix D) and is publicly available on Hugging Face.[10] The complete list of 32 figure type categories (extended from an initial eleven during manual validation), including statistics on their distribution in SciVQA are provided in Appendix E.

3.3 Evaluation metrics

Since the SciVQA dataset includes both non-binary questions, where the order of correctly predicted options can vary (e. g., A,B,C vs. C,B,A), and those requiring free-form answers, evaluation based on the exact match becomes insufficient. Therefore, we opted to use precision, recall and F1 scores of ROUGE-1, ROUGE-L (Lin, 2004), and BertScore (Zhang* et al., 2020) to capture both lexical and semantic similarity between gold references and predictions. The final ranking of the systems was determined based on the average F1 score across the three metrics. Specifically, for each system, we compute F1 scores of ROUGE-1, ROUGE-L, and BERTScore (across all questions), sum the results, and divide by the total number of metrics (i. e., three).

4 System descriptions

For the SciVQA challenge, we provide both a baseline model and human judgments to evaluate the task's difficulty and establish an upper-bound benchmark. In this section, we first outline our methodology for evaluating human performance on SciVQA. Then we describe our baseline model and the systems from five teams that submitted results to the leaderboard and corresponding reports.

Human judgments. To evaluate human performance on SciVQA, we distribute the test set across five annotators such that each receives 120 images and 840 associated QA pairs. Each student is assigned instances annotated by a different student to ensure they have not seen the questions before and have no prior knowledge of the gold answers. The task is to provide an answer given a figure image, its caption, type, and a question. Students are instructed to produce concise answers, use a template response for unanswerable questions (see §3.2), and indicate *"I don't know"* if they do not understand the question or believe no correct option is present in a multiple-choice scenario (non-binary questions). We use Label Studio configured similarly to the SciVQA human validation project (see Figure 12 in Appendix C).

Baseline. As a baseline, we use the closed-source GPT-4.1-mini model, since GPT-4 variants have demonstrated strong performance on VQA over figures (Mukhopadhyay et al., 2024; Wu et al., 2024; Wang et al., 2024b).[11] The model is run via API in a few-shot setting to enable in-context learning (Brown et al., 2020). We adopt role prompting (Schulhoff et al., 2025) to guide the model toward domain-specific reasoning, and dynamically select examples from the training set that are similar to the given test sample, as this strategy can enhance performance (Liu et al., 2022; Min et al., 2022). We

[9] https://github.com/esborisova/SciVQA/blob/main/data/SciVQA%20annotation%20guidelines.pdf

[10] https://huggingface.co/datasets/katebor/SciVQA

[11] The code for our baseline is available here: https://github.com/esborisova/SciVQA/tree/main/src/baseline

select five examples[12] matching the QA pair type and figure type of the query. If there are not enough samples with the same figure type, we randomly choose examples that share the same question type but differ in figure type. Note that for unanswerable instances, we exclude QA pair type metadata and provide two unanswerable examples along with three randomly selected samples from other question types, as including type information or using only unanswerable examples would reveal the gold answer. We define both the system prompt and the user prompt for the model. The former comprises the task instruction, examples of QA pairs, and metadata such as QA pair type (for answerable questions), figure caption, and its type (see Figure 8 in Appendix B). The user prompt includes the target question, its type (for answerable questions), an image of the target figure and its caption (see Figure 9 in Appendix B). We dynamically adjust answer format instructions based on the question type and post-process predictions to ensure they match the required structure.

ExpertNeurons. The team proposes Retrieval Augmented VQA with a Vision Language Model (RAVQA-VLM) framework (Bhat et al., 2025) which: 1. encodes images of figures and their associated metadata (caption, figure ID, type) into dense embeddings, 2. retrieves relevant context from the source scholarly papers using a dense passage retriever (Karpukhin et al., 2020), 3. and combines visual features, retrieved text, and the question as an input to an MLLM. ExpertNeurons adopts InternVL3-14B (Zhu et al., 2025) as a base model and conducts experiments using four settings. In the first, they use the vanilla version of InternVL3-14B, while in the second they fine-tune it on the SciVQA dataset using Low-Rank Adaptation (LoRa, Hu et al., 2022). The third setting additionally incorporates the RAVQA-VLM pipeline and enhances image sharpness using the Lanczos resampling technique (Turkowski, 1990; Duchon, 1979). The final approach augments the SciVQA training set with 2,500 ChartQA samples for fine-tuning InternVL3-14B.

THAii_LAB. This solution, QwenChart (Ventura et al., 2025), involves instruction fine-tuning of Qwen2.5-VL (Bai et al., 2025) models (7 and 72 billion parameters) on the SciVQA data using LoRa. THAii_LAB employs a dynamic prompting strategy with Chain-of-Thought (CoT, Wei et al., 2022) to convert each instance of SciVQA into conversation-based queries. The prompt includes task instructions, a figure image, its caption, a corresponding question, figure and question type details. Additionally, they evaluate the generalisation ability of QwenChart by testing it on out-of-domain data, namely the ChartQA benchmark.

Coling_UniA. The participants develop a system that leverages two MLLMs, InternVL3-78B and Pixtral-Large-Instruct-2411,[13] selecting the final answer based on model confidence level (Jaumann et al., 2025). The choice of model and prompting strategy is conditioned on the figure and QA pair types. For few-shot, they explore two main methods to retrieve candidate examples from the SciVQA training set: 1. using question similarity based on Sentence-BERT embeddings (Reimers and Gurevych, 2019), and 2. leveraging question and image similarity using embeddings from either CLIP (Radford et al., 2021) or BLIP-2 (Li et al., 2023). To improve MLLM configuration selection, Coling_UniA also merges rare figure types under a common category. For the experiments, they utilise the image of a figure, associated question, figure caption, and figure type labels.

florian. This team conducts a series of experiments with GPT-4o-mini and two variants of Qwen2.5-VL (7 billion and 32 billion parameters) (Schleid et al., 2025). They evaluate the performance of the models in zero- vs. one-shot setting and compare fine-tuning Qwen2.5-VL using the original SciVQA training split vs. its augmented version with additional instances from SpiQA (Pramanick et al., 2024) and ArXivQA (Li et al., 2024a). For all experiments, florian uses images of figures and their captions as an input.

Infyn. The team focuses on prompt engineering exploring the capabilities of InternVL3-8B, Qwen2.5-VL-7B-it, Bespoke-MiniChart-7B,[14] and Phi-4-multimodal (5.6 billion parameters) (Microsoft et al., 2025) models (Movva and Marupaka, 2025). Infyn designs a set of task-specific instructions for the zero-shot setting that incorporate the figure image, caption, figure type, and

[12] Due to API cost constraints, we limit the number of examples. However, including more samples could potentially lead to better results.

[13] https://huggingface.co/mistralai/Pixtral-Large-Instruct-2411

[14] https://huggingface.co/bespokelabs/Bespoke-MiniChart-7B

System	Rank	ROUGE-1			ROUGE-L			BertScore			Avg. F1
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	
Human	–	**0.8291**	**0.8347**	**0.8337**	**0.8285**	**0.8342**	**0.8330**	**0.9826**	**0.9822**	**0.9832**	**0.8801**
Baseline	–	0.7062	0.7139	0.7093	0.7055	0.7131	0.7086	0.9756	0.9762	0.9753	0.7957
ExpertNeurons	1	**0.8049**	**0.8086**	**0.8109**	**0.8043**	**0.8080**	**0.8103**	**0.9849**	**0.9850**	**0.9849**	**0.8647**
THAii_LAB	2	0.7899	0.7960	0.7949	0.7892	0.7953	0.7942	0.9839	0.9841	0.9840	0.8543
Coling_UniA	3	0.7862	0.7970	0.7860	0.7856	0.7964	0.7854	0.9817	0.9826	0.9812	0.8512
florian	4	0.7631	0.7658	0.7698	0.7621	0.7648	0.7689	0.9831	0.9830	0.9835	0.8361
Infyn	5	0.7350	0.7438	0.7437	0.7345	0.7434	0.7432	0.9787	0.9784	0.9795	0.8161
Soham Chitnis	6	0.7057	0.7190	0.7048	0.7052	0.7186	0.7043	0.9801	0.9820	0.9786	0.7970
psr123	7	0.6068	0.6089	0.6170	0.6056	0.6078	0.6156	0.9587	0.9590	0.9588	0.7237

Table 1: Evaluation results of the systems submitted to the SciVQA shared task, including human performance and baseline model. The highest scores are highlighted with grey shading and bold font. “Avg.” denotes average F1 score across ROUGE-1, ROUGE-L, and BertScore.

QA pair type, which are then combined into a single baseline prompt. They further extend this prompt by including CoT and self-reflection reasoning (Wang et al., 2025). They evaluate individual models as well as an ensemble approach, in which either Qwen2.5-VL-7B-it, Bespoke-MiniChart-7B, or Phi-4-multimodal is selected depending on the given figure type.

5 Results

Human vs. automated systems. The final results for the SciVQA challenge are presented in Table 1. The human judgments outperform automatic systems, with a maximum gap of 23%. Overall, the accuracy across individual annotators is similar: the largest difference is up to 6% in recall values and up to 3% in average F1 score across ROUGE-1, ROUGE-L, and BertScore (see Table 5 in Appendix F). This could be partially attributed to the students' prior familiarity with the task and QA pair types. The questions could also be relatively simple for humans due to their closed-ended nature and the annotators' expertise in CL. Among all the predictions, 27 questions are answered with “I don't know”, commonly due to unclear, ambiguous or incorrectly phrased questions. For instance, some questions fail to specify which subplot should be considered when an attribute is present in multiple subplots or they refer to a wrong attribute (e. g., colour, axis, value) in the graph.

Across all automatic solutions, five out of seven teams exceed our baseline. The highest scores are achieved by ExpertNeurons, using the fine-tuned InternVL3-14B model coupled with RAVQA-VLM and data augmentation. Their system surpasses our baseline by up to about 11%, while trailing behind human performance by approximately 2-3%. These findings suggest that including relevant context, along with cross-domain data, can enhance an MLLM's reasoning and generalisation abilities. QwenChart (with 7 billion parameters), proposed by THAii_LAB, ranks next. However, the team reports that their system does not generalise well to out-of-domain data, resulting in a performance drop on ChartQA. They also observe that model robustness varies depending on the question and figure type. In particular, QwenChart performs worst on infinite visual QA pairs and on figures categorised as other or containing multiple subplots with mixed types. Our baseline follows a similar trend, with visual questions, especially without predefined answer options, being more challenging for GPT-4.1-mini than non-visual ones (see Table 6 in Appendix G). THAii_LAB is closely followed by Coling_UniA, whose approach combines two MLLMs and confidence-based answer selection. The difference in scores is less than 1%. These results are interesting given that the two systems rely on different base MLLMs, prompting strategies, and learning approaches. Such a small gap highlights that while fine-tuning is effective, competitive performance can be achieved through carefully designed prompts. Similar to THAii_LAB, Coling_UniA notes that their model performs worse on infinite visual QA pair types.

Ranking fourth, florian falls behind the top three teams by about 2–5%. Their final system is based on Qwen2.5-VL (with 32 billion parameters) fine-tuned on the original SciVQA data. In line with prior observations, florian highlights that infinite visual QA pairs pose a challenge for the model. However, unlike ExpertNeurons, they find that augmenting SciVQA leads to reduced performance, although additional instances are sourced from schol-

Error type	Description
Visual attribute reasoning	Fails to correctly recognise visual attributes (e. g., colour, shape), comparing magnitudes or positions of those properties (e. g., “higher than”, “below”).
Text recognition and extraction	Fails to correctly extract labels, values, phrases, etc. This includes both cases with completely incorrect extraction and those failing to reproduce text labels, names, short phrases, exactly as they appear in the figure image or caption.
Numerical value formatting	Fails to output the correct precision (too few or too many decimal places), inconsistent/incorrect in handling of units (adding or omitting units) or representing ranges/approximate values.
Incomplete/partially correct list of items	Fails to output a complete list of expected items where all are correctly identified, e. g., for non-binary questions.
Arithmetic reasoning	Fails to correctly compute the value. This includes errors in addition, subtraction, multiplication, division, percentages, ratios or any arithmetic operation necessary for the correct response.
Other	Issues not covered by any of the five categories listed above.

Table 2: The list of error types and their definitions.

arly papers, matching the target domain. Such disparity may stem from imbalances in QA types as well as differences in format of QA pairs between SpiQA, ArXivQA and SciVQA. In this regard, figures and questions from ChartQA (used by ExpertNeurons) may be better aligned with the SciVQA dataset, especially since both include visual questions category. Infyn secures the fifth place, achieving an average F1 score of approximately 0.82 by using a model ensemble approach combined with custom prompts. Finally, Soham Chitnis and psr123 close the ranking falling behind other teams by up to about 11% and 20%, respectively. Notably, Soham Chitnis achieves scores comparable to our baseline, with a maximum difference of less than 1%. In contrast, the solution by psr123 does not surpass the SciVQA baseline, falling short by approximately 7% in average F1 score.

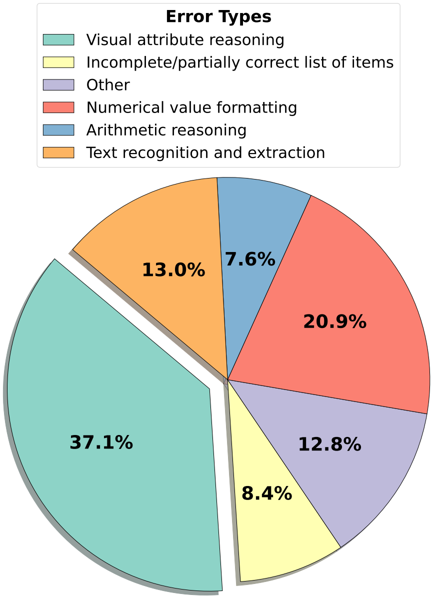


Figure 3: Distribution of error types in the predictions of the SciVQA baseline model.

Error analysis To gain insights into the common issues affecting performance, we conduct an error analysis based on the predictions from the SciVQA baseline. We identify 1,564 incorrectly answered questions based on an exact match between gold and predictions. Among those, 202 correspond to the unanswerable QA pair type, where the model simply produced an answer. To analyse the rest 1,362 cases, we generate an initial summary of errors with Gemini-2.5-Pro (see prompt in Figure 10 in Appendix B). Then we manually group those errors into the six categories listed in Table 2 and assign Google spreadsheets with 270-273 incorrectly predicted instances to five students for annotation. Additionally, we also include a “No errors” category to account for cases where the prediction is correct (e. g., gold is incorrect or incomplete).

Figure 3 shows the resulting distribution of error types, excluding the “No errors” cases. The examples of instances per error type are provided in Appendix H. The most common failures (37.1%) are associated with visual attribute reasoning. This finding, together with observations from the shared task participants, suggests that current MLLMs still struggle with interpreting visual information. Previous studies (Mukhopadhyay et al., 2024) also report challenges in MLLMs’ visual reasoning such as errors associated with colour encoding, especially when it comes to similar shades. The second largest group of errors (20.9%) is related to numerical value formatting. The most frequent mismatches involve the absence of approximations or ranges and slight numerical discrepancies. This indicates that GPT-4.1-mini may not have fully learned the expected answer formatting from the given examples. Text recognition and extraction along with the other errors account for 13% and 12.8%, respectively. The former often includes failures in reproducing the required formatting of text (e. g., see Figure 19). For the "Other" category, we observe that annotators specify cases where either the gold answer is incorrect or both the gold answer and the prediction are valid. Similarly, several such cases appear under the "No errors" label. In total, 111 out of 4200 gold instances are flagged as being incorrect. Given the large scale of the dataset and the error-prone nature of manual annotation (Klie et al., 2024), one round of human validation of synthetic QA pairs may have been insufficient, resulting in some noise. Although the percentage of annotation errors is rather small (approximately 2.6%), they likely affected the final evaluation scores. Notably, the "Incomplete/partially correct list of items" category constitutes only 8% of all errors, followed by arithmetic reasoning failures (7.6%).

6 Conclusion

In this paper, we presented an overview of the first SciVQA shared task. The challenge attracted seven submissions, five of which outperformed our baseline. The results reveal that, while automated systems can achieve strong performance on the newly proposed SciVQA benchmark, they remain behind human judgments. Furthermore, the findings indicate that fine-tuning on cross-domain data, combined with relevant contextual information from source papers, leads to the best results. However, domain adaptation and data augmentation is not always required, and carefully designed prompting strategies can achieve very close results (about 2% gap). Additionally, we observe that current MLLMs struggle most with visual reasoning, as their accuracy drops on QA pairs addressing visual attributes of figures.

Limitations

Although this study sheds light on the abilities of current MLLMs to reason over scientific figures, it is not without limitations. First, the evaluation relies on automated metrics, ROUGE and BertScore, which may fall short when handling free-form answers. BertScore is also less suitable for non-binary questions, since answer options are short, leading to high similarity scores being assigned to distinct choices (e. g., A vs. B). Additional manual review could be beneficial for the analysis of prediction quality. Second, SciVQA provides a single gold reference, whereas multiple valid answers may exist. Extending the dataset to include several references could improve the fairness of the evaluation process. Third, the SciVQA test set contains a few annotation errors which can influence scoring. As a next step, we plan another manual revision to correct these errors and improve data quality. Finally, this study focuses solely on closed-ended QA in English, and we leave the extension of SciVQA to open-ended multilingual QA for future work.

Acknowledgments

The work received funding through the German Research Foundation (DFG) project NFDI for Data Science and Artificial Intelligence, NFDI4DS [15] (no. 460234259). We would like to thank Valentina Tretti, Radovan Milovic, Ardalan Khazraei, David Raul Carranza Navarrete, Shane John Paul Newton, Melina Plakidis, Emma Carballal, and Maria Francis for annotating the SciVQA data, error types, and conducting human evaluation. We also thank Raia Abu Ahmad for the fruitful discussions related to the dataset construction and competition set up.

Ethics statement

SciVQA does not contain any sensitive or personal data. The images of the figures used to construct the SciVQA benchmark are sourced from the publicly available datasets. We comply with their respective licenses and usage terms. The annotators were compensated according to a standard payment scheme and were informed about the intended use of their annotations.

References

Saleem Ahmed, Bhavin Jawade, Shubham Pandey, Srirangaraj Setlur, and Venu Govindaraju. 2023. Realcqa: Scientific chart question answering as a testbed for first-order logic. In *Document Analysis and Recognition - ICDAR 2023*, pages 66–83, Cham. Springer Nature Switzerland.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

[15] 19https://www.nfdi4datascience.de

Nagaraj Bhat, Joydeb Mondal, and Srijon Sarkar. 2025. ExpertNeurons at SciVQA-2025: Retrieval augmented VQA with vision language model (RAVQA-VLM). In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2019. LEAF-QA: Locate, encode & attend for figure question answering. *Preprint*, arXiv:1907.12861.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 143–152.

Claude Duchon. 1979. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology - J APPL METEOROL*, 18:1016–1022.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. ChartLlama: A multimodal LLM for chart understanding and generation. *Preprint*, arXiv:2311.16483.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *Preprint*, arXiv:2403.12027.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA. Association for Computational Linguistics.

Christian Jaumann, Annemarie Friedrich, and Rainer Lienhart. 2025. Coling-UniA at SciVQA 2025: Few-shot example retrieval and confidence-informed ensembling for multimodal large language models. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria. Association for Computational Linguistics.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

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*, pages 11817–11837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. ACL-Fig: A dataset for scientific figure classification. *Preprint*, arXiv:2301.12293.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Wonjoong Kim, Sangwu Park, Yeonjun In, Seokwon Han, and Chanyoung Park. 2024. SIMPLOT: Enhancing chart question answering by distilling essentials. *Preprint*, arXiv:2405.00021.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Shengzhi Li and Nima Tajbakhsh. 2023. SciGraphQA: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *Preprint*, arXiv:2308.03349.

Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. 2024b. Synthesize step-by-step: Tools, templates and LLMs as data generators for reasoning-based chart VQA. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13613–13623, Los Alamitos, CA, USA. IEEE Computer Society.

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.

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

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023b. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

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)*, pages 1287–1310, Mexico City, Mexico. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

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*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025a. ChartQAPro: A more diverse and challenging benchmark for chart question answering. *Preprint*, arXiv:2504.05506.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore. Association for Computational Linguistics.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2025b. ChartGemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643, Abu Dhabi, UAE. Association for Computational Linguistics.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7775–7803, Bangkok, Thailand. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over scientific plots. *Preprint*, arXiv:1909.00997.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prahitha Movva and Naga Harshita Marupaka. 2025. Enhancing scientific visual question answering through multimodal reasoning and ensemble modeling. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria. Association for Computational Linguistics.

Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024.

Unraveling the truth: Do VLMs really understand charts? a deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16696–16717, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*, volume 37, pages 118807–118833. Curran Associates, Inc.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. SciFIBench: Benchmarking large multimodal models for scientific figure interpretation. In *Advances in Neural Information Processing Systems*, volume 37, pages 18695–18728. Curran Associates, Inc.

Florian Schleid, Jan Strich, and Chris Biemann. 2025. Visual question answering on scientific charts using fine-tuned vision-language models. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria. Association for Computational Linguistics.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

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

Ken Turkowski. 1990. *Filters for common resampling tasks*, page 147–165. Academic Press Professional, Inc., USA.

Viviana Ventura, Lukas Kleybolte, and Alessandra Zarcone. 2025. Instruction-tuned QwenChart for chart question answering. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria. Association for Computational Linguistics.

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. 2024a. MinerU: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Preprint*, arXiv:2504.08837.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jingxuan Wei, Nan Xu, Guiyong Chang, Yin Luo, BiHui Yu, and Ruifeng Guo. 2024. mChartQA: A universal benchmark for multimodal chart question answer based on vision-language alignment and reasoning. *Preprint*, arXiv:2404.01548.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40. Language in Vision.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. ChartInsights: Evaluating multimodal large language models for low-level chart question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200, Miami, Florida, USA. Association for Computational Linguistics.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. 2024. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *Preprint*, arXiv:2407.20174.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1898, Miami, Florida, USA. 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*.

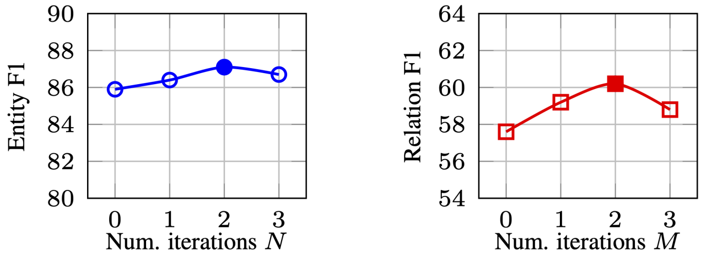
Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding via visual language pretraining on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1314–1326, Toronto, Canada. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Question answering pair types

Question answering pair type	Definition
Closed-ended	It is possible to answer a question based only on a given data source, i. e., a figure image and/or its caption. No additional resources such as the main text of a publication, other documents, figures or tables are required.
Unanswerable	It is not possible to infer an answer based solely on a given data source.
Infinite answer set	There are no predefined answer options.
Finite answer set	There is a limited range of answer options.
Binary	Requires a yes/no or true/false answer.
Non-binary	Requires to choose from a set of (four) predefined answer options where one or more are correct.
Visual	Addresses or incorporates information on one or more of the six visual attributes of a figure, i. e., shape, size, position, height, direction or colour.
Non-visual	Does not involve any of the six pre-defined visual aspects of a figure.

Table 3: The list of question answering pair types and their definitions.



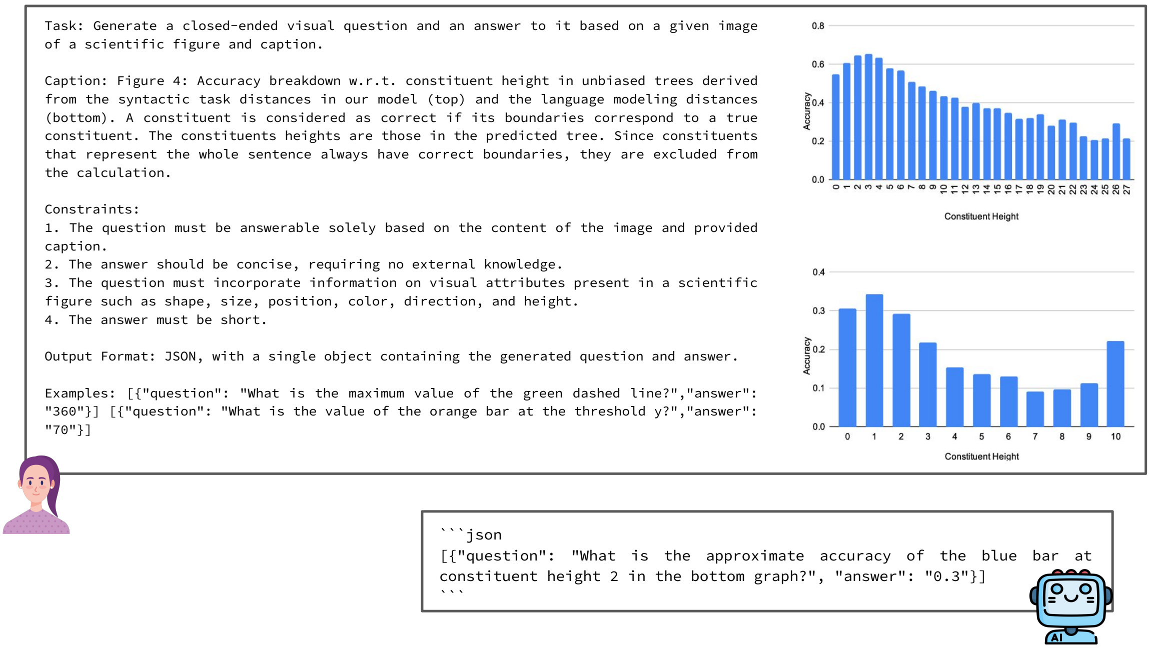
(a) Entity F1 with different number of `CorefProp` iterations N.

(b) Relation F1 with different number of `RelProp` iterations M.

```
Figure Caption: Figure 3: F1 score of each layer on ACE development set for different number
of iterations. N = 0 or M = 0 indicates no propagation is made for the layer.

Question Type: Closed-ended infinite answer set visual
Question: What is the F1 score of the red line at M = 2?
Answer: 60

Question Type: Closed-ended infinite answer set non-visual
Question: What is the F1 score for the entity layer when there are no iterations?
Answer: 86

Question Type: Closed-ended finite answer set binary visual
Question: Is the highest F1 score for Entity in the left plot achieved at N=2?
Answer: Yes

Question Type: Closed-ended finite answer set binary non-visual
Question: Does the number of iterations impact the F1 score for both Entity and Relation?
Answer: Yes

Question Type: Closed-ended finite answer set non-binary visual
Question: Which graph shows the F1 score for RelProp iterations?
Answer options: A: The graph on the left B: The graph on the right C: Both graphs D: Neither
graph
Answer: B

Question Type: Closed-ended finite answer set non-binary non-visual
Question: Which kind of F1 is above 75% for all iterations?
Answer options: A: Entity B: Relation C: Both D: Neither
Answer: A

Question Type: Unanswerable
Question: What is the F1 score of the entity layer after 2 iterations of propagation with N =
1 and M = 2?
Answer: It is not possible to answer this question based only on the provided data.
```

Figure 4: Example of a figure and seven question answering pair types associated with it. The sample is taken from the SciVQA training set.

B Prompts



Task: Generate a closed-ended visual question and an answer to it based on a given image of a scientific figure and caption.

Caption: Figure 4: Accuracy breakdown w.r.t. constituent height in unbiased trees derived from the syntactic task distances in our model (top) and the language modeling distances (bottom). A constituent is considered as correct if its boundaries correspond to a true constituent. The constituents heights are those in the predicted tree. Since constituents that represent the whole sentence always have correct boundaries, they are excluded from the calculation.

Constraints:
1. The question must be answerable solely based on the content of the image and provided caption.
2. The answer should be concise, requiring no external knowledge.
3. The question must incorporate information on visual attributes present in a scientific figure such as shape, size, position, color, direction, and height.
4. The answer must be short.

Output Format: JSON, with a single object containing the generated question and answer.

Examples: [{"question": "What is the maximum value of the green dashed line?","answer": "360"}] [{"question": "What is the value of the orange bar at the threshold y?","answer": "70"}]

```json
[{"question": "What is the approximate accuracy of the blue bar at constituent height 2 in the bottom graph?", "answer": "0.3"}]
```

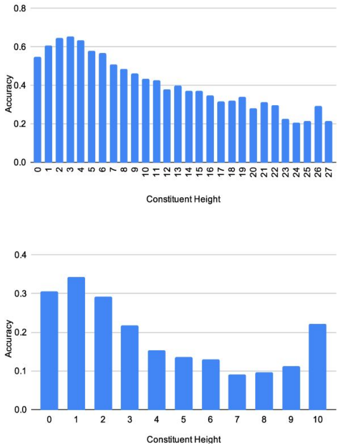
Figure 5: Example of a prompt for generating a closed-ended visual question with infinite answer set using Gemini-1.5-Flash.

Task: You are given an image of a figure extracted from a scholarly paper and its caption. Identify whether this image contains a compound or non-compound figure. Non-compound means that there is only one figure object in an image. Compound means there are two or more figure objects in an image. If a figure is compound, determine the number of subfigures.

Caption: Figure 4: Accuracy breakdown w.r.t. constituent height in unbiased trees derived from the syntactic task distances in our model (top) and the language modeling distances (bottom). A constituent is considered as correct if its boundaries correspond to a true constituent. The constituents heights are those in the predicted tree. Since constituents that represent the whole sentence always have correct boundaries, they are excluded from the calculation.

Output Format: JSON containing the figure type.

Examples: [{"compound": "True", "subfigures": "6"}], [{"compound": "False", "subfigures": "0"}].



~~~
```json
[{"compound": "True", "subfigures": "2"}]
```
~~~

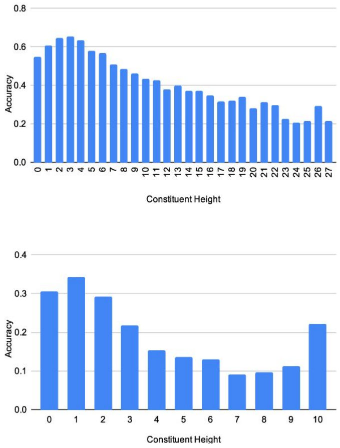
Figure 6: Example of a prompt for classifying figures into compound and non-compound using Gemini-1.5-Flash.

Task: You are given an image of a figure and its caption extracted from a scholarly paper. Classify this figure into one of the following types: bar chart, box plot, confusion matrix, line chart, pie chart, scatter plot, pareto chart, venn diagram, architecture diagram, neural networks, tree.

Caption: Figure 4: Accuracy breakdown w.r.t. constituent height in unbiased trees derived from the syntactic task distances in our model (top) and the language modeling distances (bottom). A constituent is considered as correct if its boundaries correspond to a true constituent. The constituents heights are those in the predicted tree. Since constituents that represent the whole sentence always have correct boundaries, they are excluded from the calculation.

Output format: JSON, with a single object containing the figure type.

Example: [{"type": ""}].



~~~
```json
[{"type": "bar chart"}]
```
~~~

Figure 7: Example of a prompt for classifying figures into types using Gemini-1.5-Flash.

```
You are an expert scientific figure analyst specializing in academic publications.
Your task is to answer questions about scientific figures and their captions accurately and concisely.
Answer the given question based *solely* on the information visible in the figure and its provided caption.

The user message will include a ‘Question Type’. Adhere strictly to the following rules for formatting your response
based on the question type:

- For ‘closed-ended finite answer set binary visual’ or ’closed-ended finite answer set binary non-visual’:
- Respond ONLY with ‘Yes’ or ‘No’.
- Do NOT add any other text, explanations, or punctuation.
- Your entire response must be exactly one word: either ‘Yes’ or ‘No’.

- For ‘closed-ended finite answer set non-binary visual’ or ’closed-ended finite answer set non-binary non-visual’:
- Identify the correct option(s) from the provided ‘Answer Options’.
- Respond ONLY with the letter(s) of the correct option(s) as listed.
- For a single correct option, provide only its letter (e.g., ‘B’).
- For multiple correct options, list ALL correct letters separated by commas with NO SPACES (e.g., ‘A,C,D’).
- Ensure ALL correct options are listed and NO incorrect ones.
- Do NOT add any other text, explanations, or surrounding punctuation.

- For ‘closed-ended infinite answer set visual’ or ’closed-ended infinite answer set non-visual’:
- Provide a brief, direct answer.
- This answer must be a value, a short phrase, a specific name, a label, or a list of values read directly from the figure
or caption.
- **For numerical values:** Read values as precisely as possible from the graph axes, data points, or labels. Include
units ONLY if they appear in the figure.
- **For non-numerical values:** Reproduce them EXACTLY as they appear in the figure or caption.
- Do NOT add any introductory phrases, explanations, or surrounding text.

- For ‘unanswerable’:
- Respond ONLY with the exact phrase: ‘It is not possible to answer this question based only on the provided data.’
- Do NOT add any other text.

IMPORTANT: Your response should ONLY contain the answer in the correct format as specified above - nothing else.
Do NOT include any additional text, explanations, comments, or contextual information.
Your answer must be based solely on the information visible in the figure and its provided caption.

Below are examples of questions and answers similar to what you will receive. Study these examples carefully to understand
the expected answer format. Your question will be in the user message after these examples:

Example 1:
Figure Caption: {caption}
Figure Type: {figure type}
Question Type: {question type}
Question: {question}
Correct answer: {answer}

Example 2:
Figure Caption: {caption}
Figure Type: {figure type}
Question Type: {question type}
Question: {question}
Correct answer: {answer}

Example 3:
Figure Caption: {caption}
Figure Type: {figure type}
Question Type: {question type}
Question: {question}
Correct answer: {answer}

Example 4:
Figure Caption: {caption}
Figure Type: {figure type}
Question Type: {question type}
Question: {question}
Correct answer: {answer}

Example 5:
Figure Caption: {caption}
Figure Type: {figure type}
Question Type: {question type}
Question: {question}
Correct answer: {answer}

REMEMBER: {answer format instruction}.
```

Figure 8: System prompt used for the SciVQA baseline model, GPT-4.1-mini. For unanswerable questions, the metadata on their type is excluded since it directly reveals the answer.

```
Figure Caption: {caption}
Figure Type: {figure type}
Question Type: {question type}
Question: {question}
```

Figure 9: User prompt used for the SciVQA baseline model, GPT-4.1-mini. For unanswerable questions, the metadata on their type is excluded since it directly reveals the answer.

```
Analyse the incorrectly predicted answers and try to find common patterns.
Here are the evaluation scores for the predictions: {scores}
Here is the JSON string with the gold and predicted answers: {JSON string}
```

Figure 10: Prompt used for Gemini-2.5-Pro to summarise the common errors in the predictions from the SciVQA baseline. JSON string contains instance IDs, questions, gold answers, predictions, information on figure and question types.

C Label Studio configuration examples

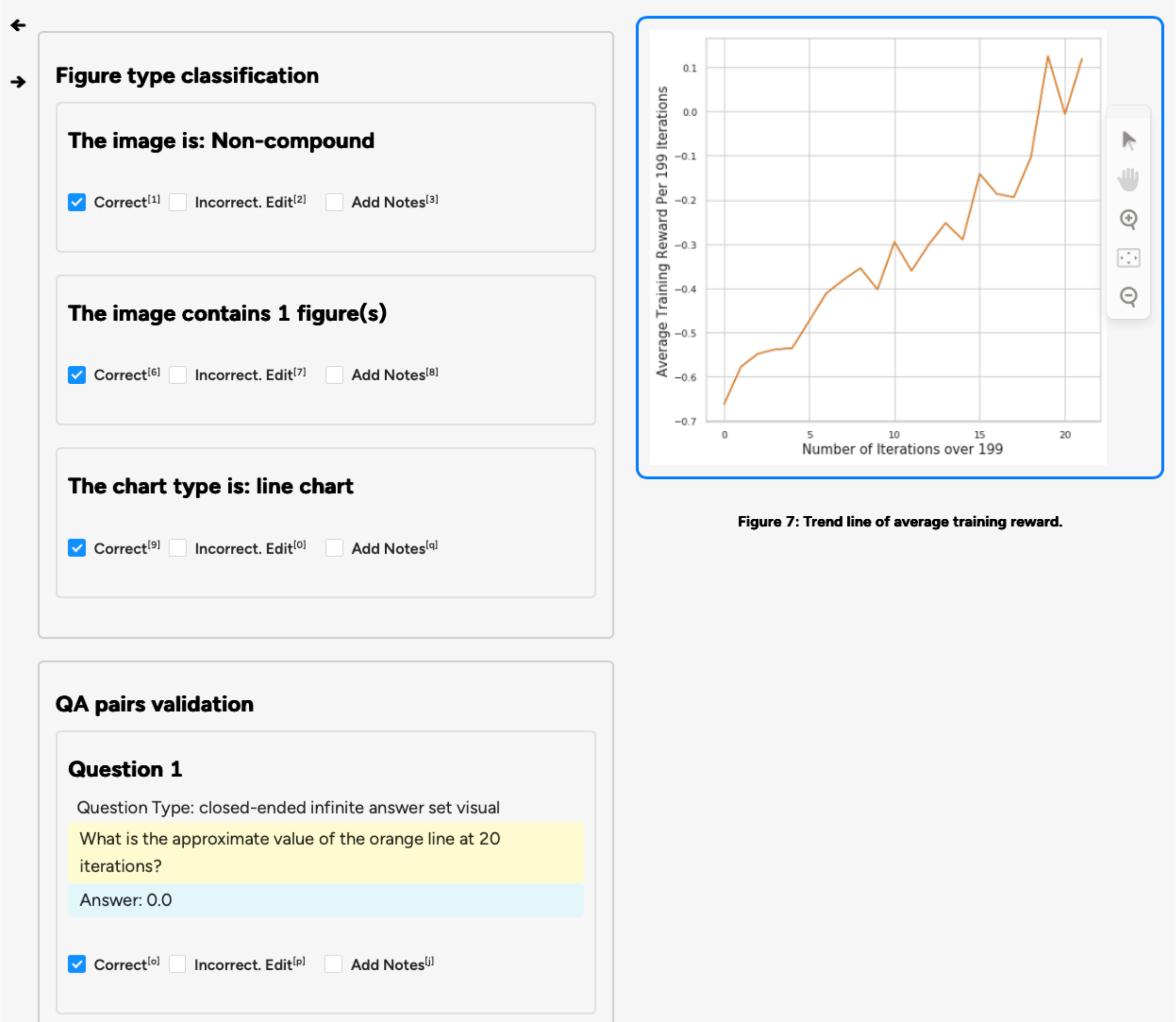


Figure type classification

The image is: Non-compound

☑ Correct[1] □ Incorrect. Edit[2] □ Add Notes[3]

The image contains 1 figure(s)

☑ Correct[6] □ Incorrect. Edit[7] □ Add Notes[8]

The chart type is: line chart

☑ Correct[9] □ Incorrect. Edit[0] □ Add Notes[q]

QA pairs validation

Question 1

Question Type: closed-ended infinite answer set visual

What is the approximate value of the orange line at 20 iterations?

Answer: 0.0

☑ Correct[o] □ Incorrect. Edit[p] □ Add Notes[j]

Figure 7: Trend line of average training reward.

Average Training Reward Per 199 Iterations
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
0
5
10
15
20
Number of Iterations over 199

Figure 11: Example setup for the human validation phase in Label Studio (a snapshot).

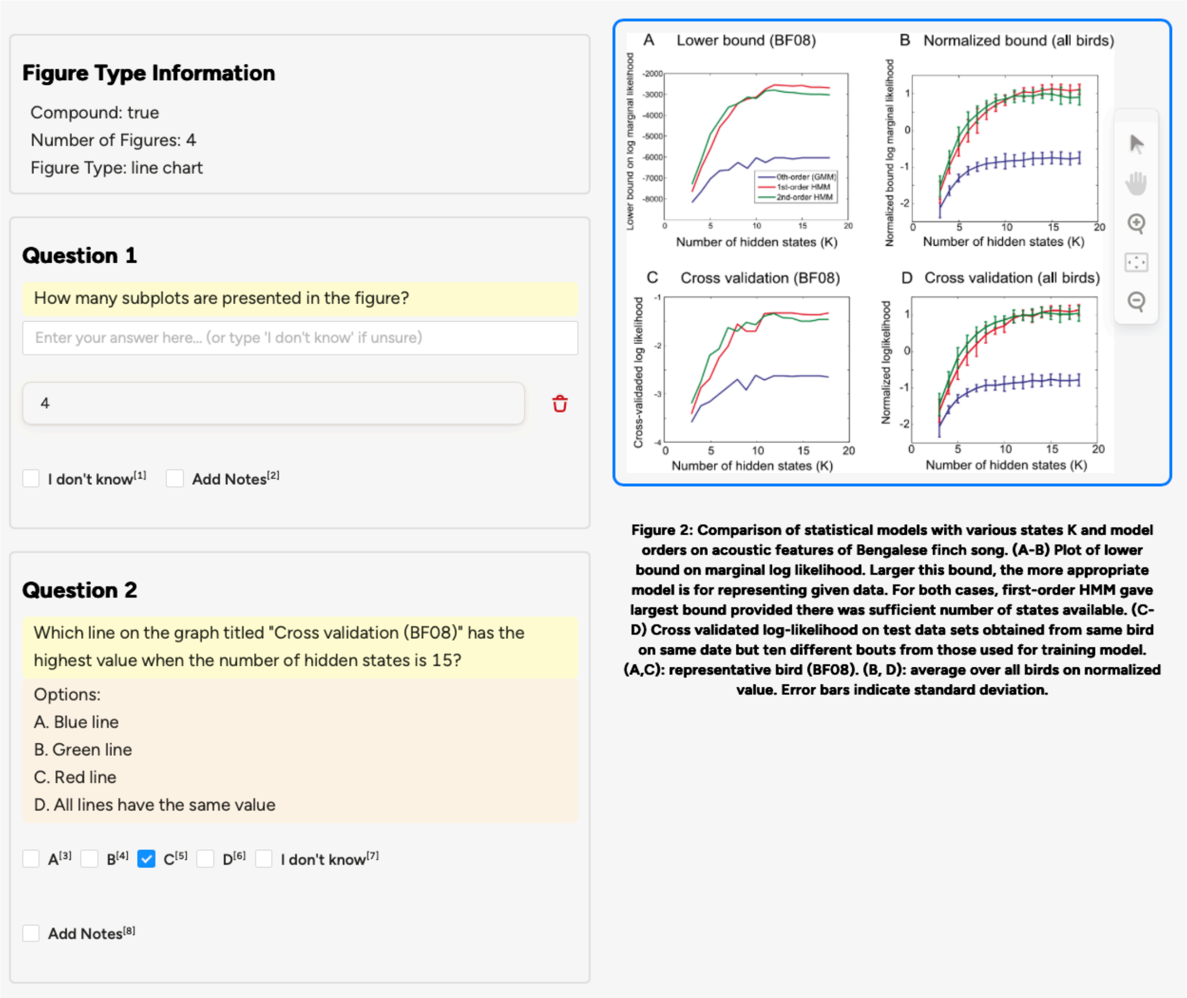

A Lower bound (BF08)
Lower bound on log marginal likelihood
-2000
-3000
-4000
-5000
-6000
-7000
-8000
0th-order (GMM)
1st-order HMM
2nd-order HMM
Number of hidden states (K)
B Normalized bound (all birds)
Normalized bound log marginal likelihood
Number of hidden states (K)
C Cross validation (BF08)
Cross-validated log likelihood
Number of hidden states (K)
D Cross validation (all birds)
Normalized loglikelihood
Number of hidden states (K)
0
5
10
15
20

Figure Type Information

Compound: true
Number of Figures: 4
Figure Type: line chart

Question 1

How many subplots are presented in the figure?

Enter your answer here... (or type 'I don't know' if unsure)

4

☐ I don't know[1] ☐ Add Notes[2]

Question 2

Which line on the graph titled "Cross validation (BF08)" has the highest value when the number of hidden states is 15?

Options:
A. Blue line
B. Green line
C. Red line
D. All lines have the same value

☐ A[3] ☐ B[4] ☑ C[5] ☐ D[6] ☐ I don't know[7]

☐ Add Notes[8]

Figure 2: Comparison of statistical models with various states K and model orders on acoustic features of Bengalese finch song. (A-B) Plot of lower bound on marginal log likelihood. Larger this bound, the more appropriate model is for representing given data. For both cases, first-order HMM gave largest bound provided there was sufficient number of states available. (C-D) Cross validated log-likelihood on test data sets obtained from same bird on same date but ten different bouts from those used for training model. (A,C): representative bird (BF08). (B, D): average over all birds on normalized value. Error bars indicate standard deviation.

Figure 12: Example setup for the human performance evaluation in Label Studio (a snapshot).

D Data distribution in SciVQA

Split	Images	QA pairs
Train	2160	15120
Validation	240	1680
Test	600	4200
Total	**3000**	**21000**

Table 4: Distribution of figure images and QA pairs in SciVQA dataset across train, validation, and test splits.

E Figure types in SciVQA

The final list of figure types based on the stylistic features comprises 32 classes: *line chart*, *bar chart*, *box plot*, *confusion matrix*, *pie chart*, *scatter plot*, *pareto chart*, *venn diagram*, *architecture diagram*, *neural networks*, *tree*, *graph*, *other*, *histogram*, *heat map*, *illustrative diagram*, *flow chart*, *violin plot*, *vector plot*, *density plot*, *faceted dot plot*, *t-sne plot*, *word-alignment grid*, *tree set*, *target plot*, *bar chart with error*, *lex plot*, *contour*, *dendogram*, *speech balloons*, *surface plot*, and *parallel coordinates plot*. As can be seen from Figure 13, line charts are the most common overall.

Figures 14 shows that the majority of the figures in SciVQA are non-compound (60.53%). Compound figures constitute 39.47% of the dataset, with those containing two sub-figures being the most prevalent (see Figure 15).

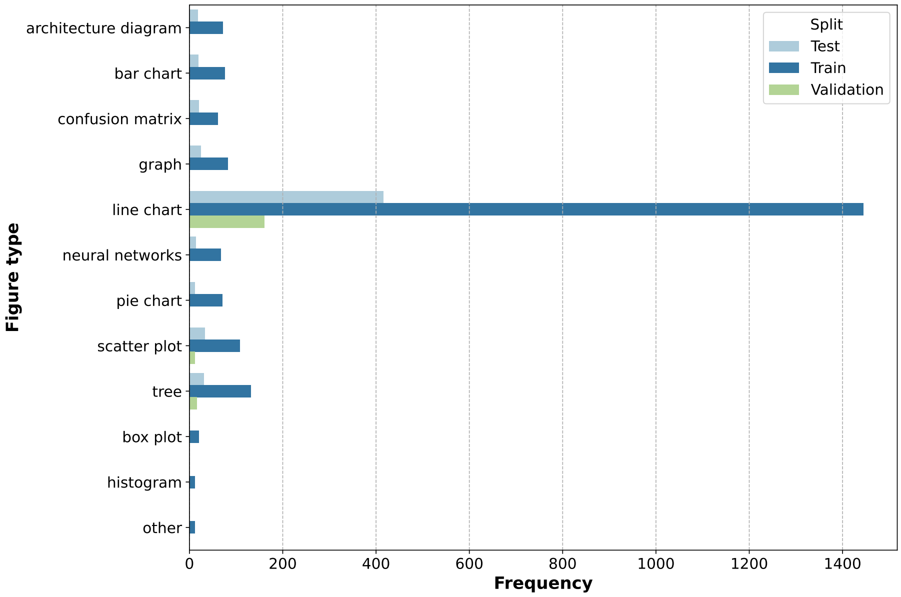


Figure 13: Distribution of figure types across train, validation, and test splits in the SciVQA dataset. Given the large number of classes (32), only those with the frequency of occurrence larger than 10 are shown.

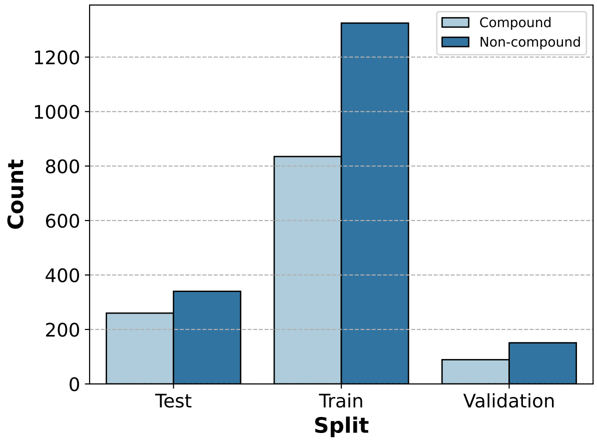


Figure 14: Distribution of compound and non-compound figures across train, validation, and test splits in the SciVQA dataset.

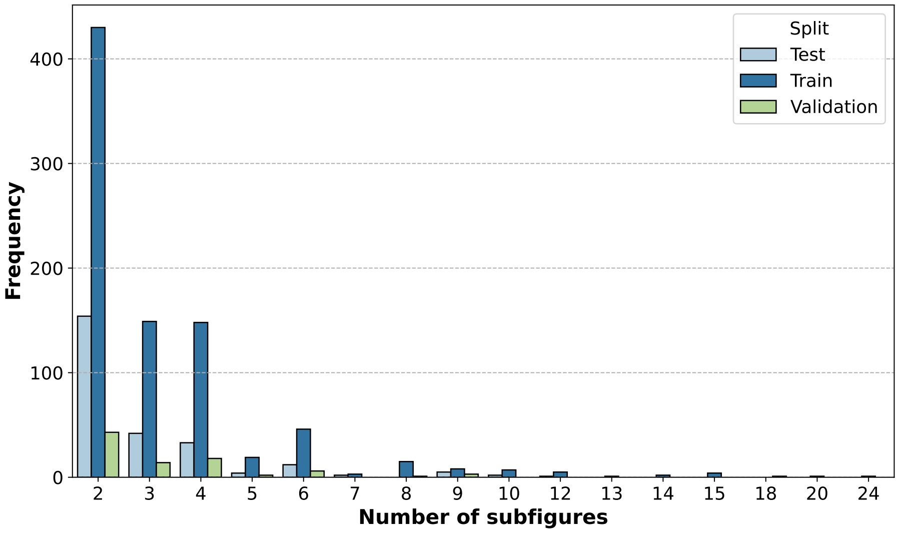


Figure 15: Distribution of the number of sub-figures in compound figures across train, validation, and test splits in the SciVQA dataset.

F Human performance

Annotator	ROUGE-1			ROUGE-L			BertScore			Avg. F1
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	
`Annotator_1`	0.8478	0.8437	0.8617	0.8464	0.8423	0.8603	0.9807	0.9791	0.9825	**0.8916**
`Annotator_2`	0.8420	0.8441	0.8501	0.8417	0.8439	0.8495	0.9809	0.9807	0.9813	**0.8882**
`Annotator_3`	0.8262	0.8322	0.8264	0.8256	0.8316	0.8258	0.9856	0.9860	0.9856	**0.8791**
`Annotator_4`	0.8218	0.8367	0.8225	0.8218	0.8367	0.8225	0.9799	0.9793	0.9807	**0.8745**
`Annotator_5`	0.8078	0.8170	0.8077	0.8073	0.8165	0.8070	0.9857	0.9859	0.9858	**0.8669**

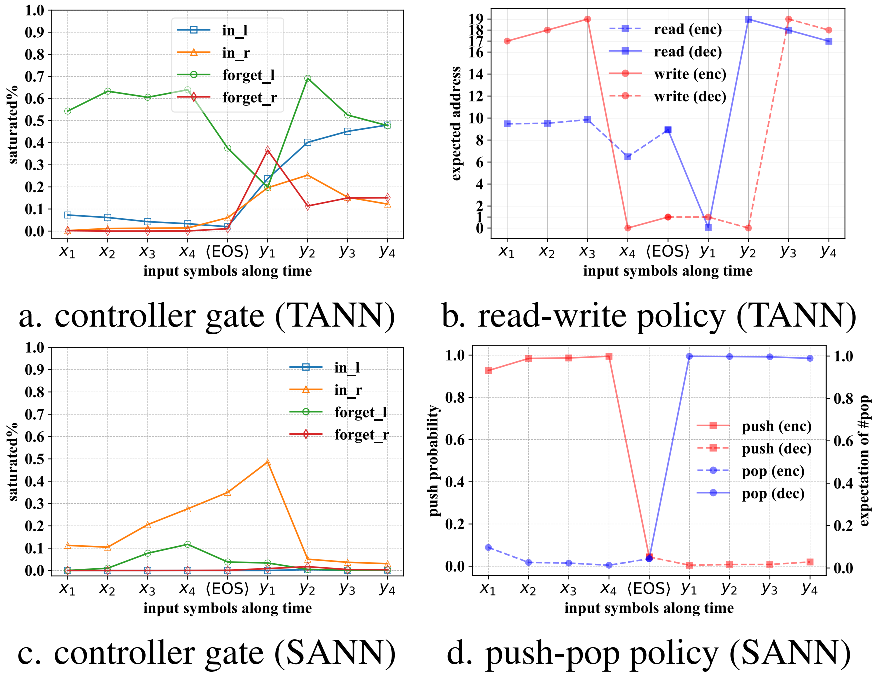
Table 5: Evaluation results of the human performance on SciVQA for each annotator. "Avg." denotes average F1 score across ROUGE-1, ROUGE-L, and BertScore.

G Baseline performance

QA pair type	ROUGE-1			ROUGE-L			BERTScore			Avg. F1
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	
finite answer set binary non-visual	0.8317	0.8317	0.8317	0.8317	0.8317	0.8317	0.9999	0.9999	0.9999	**0.8877**
finite answer set binary visual	0.7683	0.7683	0.7683	0.7683	0.7683	0.7683	0.9998	0.9998	0.9998	**0.8455**
finite answer set non-binary non-visual	0.7316	0.7276	0.7571	0.7312	0.7272	0.7567	0.9814	0.9782	0.9853	**0.8148**
finite answer set non-binary visual	0.7089	0.7124	0.7143	0.7080	0.7115	0.7135	0.9921	0.9915	0.9929	**0.8030**
infinite answer set non-visual	0.7009	0.7211	0.6998	0.6985	0.7183	0.6977	0.9623	0.9646	0.9606	**0.7872**
infinite answer set visual	0.5329	0.5673	0.5237	0.5319	0.5659	0.5229	0.9524	0.9584	0.9470	**0.6724**
unanswerable	0.6689	0.6691	0.6700	0.6687	0.6689	0.6696	0.9412	0.9406	0.9420	**0.7596**

Table 6: Evaluation results of the SciVQA baseline model across different question answering (QA) pair types. "Avg." denotes average F1 score across ROUGE-1, ROUGE-L, and BertScore.

H Examples of errors



a. controller gate (TANN)   b. read-write policy (TANN)

c. controller gate (SANN)   d. push-pop policy (SANN)

Figure Caption: Figure 3: Averaged visualization about (a and c) controller gate and (b and d) read-write policy for TANN and SANN on mirror task. Note that all the plots are derived from being averaging over 500 random samples. The x-axis shows each time step represented by input xi or output yi. The ⟨EOS⟩ represent the input delimiter.

Figure Type: line chart

Question Type: closed-ended finite answer set binary non-visual

Question: Does ‘in_r’ always have a higher saturated percentage compared to others in plot ‘c’?

Gold answer: Yes

Predicted answer: No

Figure 16: An example of an incorrect prediction by the SciVQA baseline, categorised as containing visual attribute and arithmetic reasoning errors.

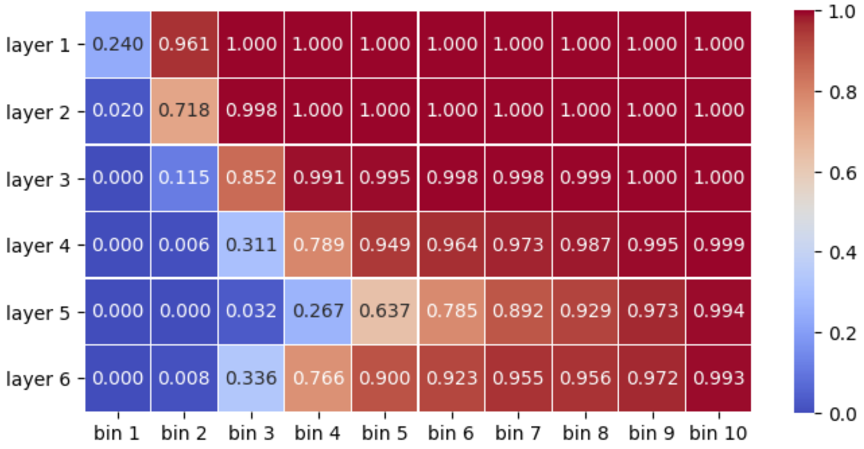


Figure Caption: Figure 9: Frequency-based classification accuracy on states from the ENDE encoder + lexical shortcuts.
Figure Type: heat map
Question Type: closed-ended infinite answer set visual
Question: What is the colour of the cell in the heatmap that is in the same row as ‘layer 2’ and the same column as ‘bin 3’?

Gold answer: Red
Predicted answer: light orange

Figure 17: An example of an incorrect prediction by the SciVQA baseline, categorised as containing visual attribute reasoning error.

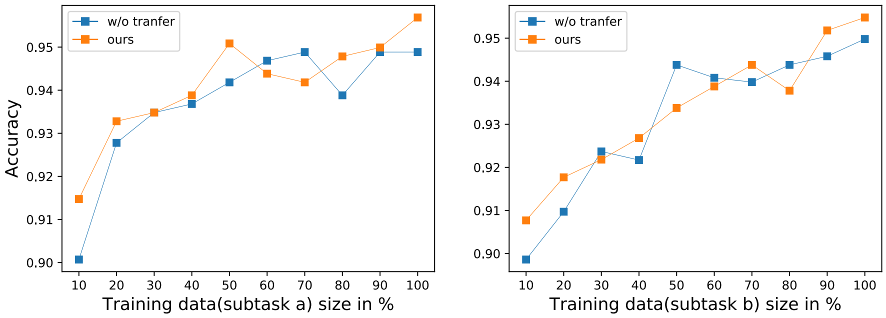


Figure Caption: Figure 4: Learning curve on the training dataset.
Figure Type: line chart
Question Type: closed-ended finite answer set non-binary non-visual
Question: Which of the following subtasks reach a value more than 0.93 at 20% training data for 'ours' methods?
Answer options: A: subtask a | B: subtask b | C: subtask c | D: All of the above

Gold answer: A
Predicted answer: A,B

Figure 18: An example of an incorrect prediction by the SciVQA baseline, categorised as containing incomplete/-partially correct list of items error.

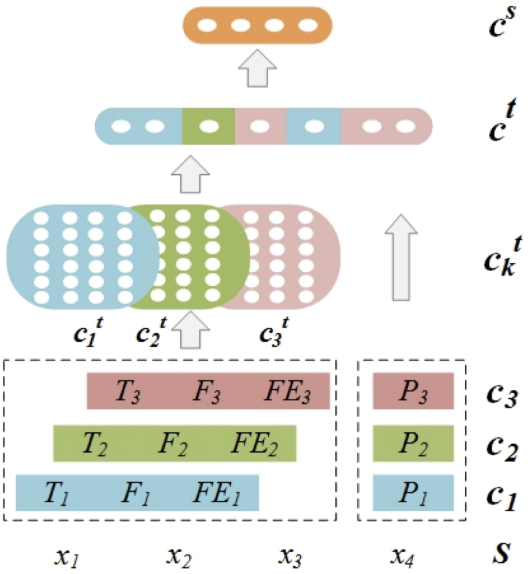


Figure Caption: Figure 4: Frame Integration Representation Model.
Figure Type: neural networks
Question Type: closed-ended infinite answer set non-visual
Question: What is the final vector c^s derived from?

Gold answer: c^t
Predicted answer: ct

Figure 19: An example of an incorrect prediction by the SciVQA baseline, categorised as containing text recognition and extraction error.

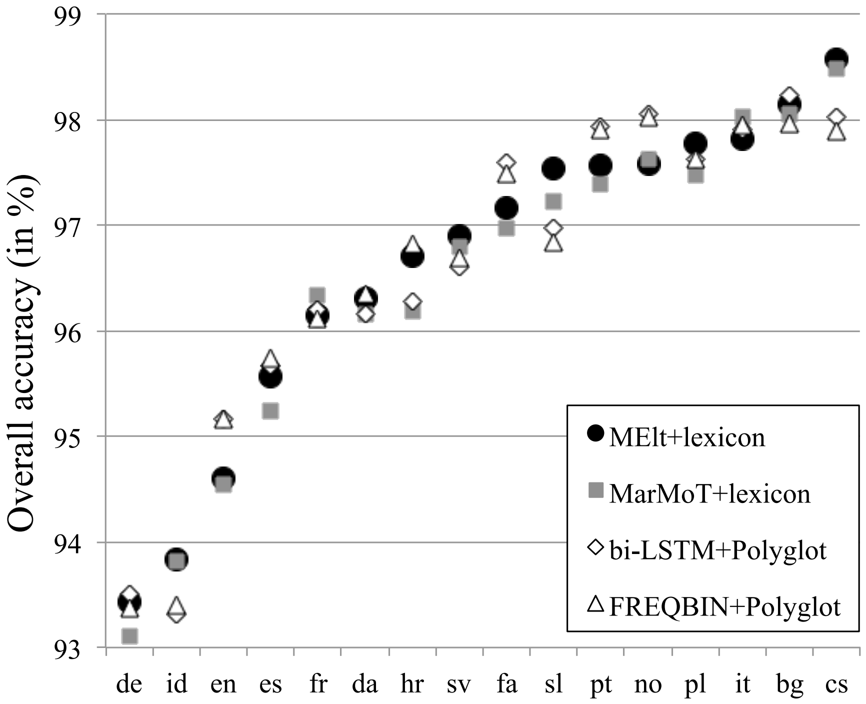


Figure Caption: Figure 1: Graphical visualisation of the overall tagging accuracies for all four types of enriched models. Detailed results are given in Table 4. Languages are sorted by increasing MElt's overall tagging scores.
Figure Type: scatter plot
Question Type: closed-ended infinite answer set visual
Question: What is the overall accuracy of the black circle marker for the language 'es'?

Gold answer: between 95 and 96
Predicted answer: 95.7

Figure 20: An example of an incorrect prediction by the SciVQA baseline, categorised as containing numerical value formatting error.