Coling-UniA at SciVQA 2025: few-shot example retrieval and confidence-informed ensembling for Multimodal Large Language Models

Christian Jaumann, Annemarie Friedrich, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Jaumann, Christian, 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 Fifth Workshop on Scholarly Document Processing (SDP 2025), 31 July 2025, Vienna, Austria*, edited by Tirthankar Ghosal, Philipp Mayr, Amanpreet Singh, Aakanksha Naik, Georg Rehm, Dayne Freitag, Dan Li, Sonja Schimmler, and Anita De Waard, 230–39. Stroudsburg, PA: Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2025.sdp-1.21.

Coling-UniA at SciVQA 2025: Few-Shot Example Retrieval and Confidence-Informed Ensembling for Multimodal Large Language Models

Christian Jaumann **Annemarie Friedrich** **Rainer Lienhart**
University of Augsburg, Germany
{firstname.lastname}@uni-a.de

Abstract

This paper describes our system for the SciVQA 2025 Shared Task on Scientific Visual Question Answering. Our system employs an ensemble of two Multimodal Large Language Models and various few-shot example retrieval strategies. The model and few-shot setting are selected based on the figure and question type. We also select answers based on the models' confidence levels. On the blind test data, our system ranks third out of seven with an average F1 score of 85.12 across ROUGE-1, ROUGE-L, and BERTS. Our code is publicly available.[1]

1 Introduction

Visual Question Answering (VQA) requires systems to answer natural language questions about visual content. The complexity of these questions can range from binary questions to free-form and open-ended questions. Existing VQA datasets address various types of images, e.g., VQA v2 focuses on real-world photos (Goyal et al., 2017), DocVQA focuses on scanned documents (Mathew et al., 2021), while ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020) focus on charts.

In this paper, we describe our system submission for the 2025 Shared Task on Scientific Visual Question Answering (SciVQA) (Borisova et al., 2025). The dataset[2] comprises 3000 real-world scientific figure images, which were collected from the ACL-Fig (Karishma et al., 2023) and SciGraphQA (Li and Tajbakhsh, 2023).

Most existing VQA approaches that focus on charts rely on models explicitly tuned for this domain (Liu et al., 2023; Han et al., 2023; Xia et al., 2024; Zhang et al., 2024). In contrast, our approach uses Multimodal Large Language Models (MLLMs) in a zero/few-shot setting without any fine-tuning. We test several strategies for retrieving few-shot examples from the training set based on question or question-and-image similarity. We find that performance varies widely by question/figure type and by MLLM. Our best-performing approach first selects highly confident answers from a configuration of an MLLM and a few-shot setting. For all remaining instances, the system configuration is varied by the instance's question type. In the official evaluation, our system ranks third.

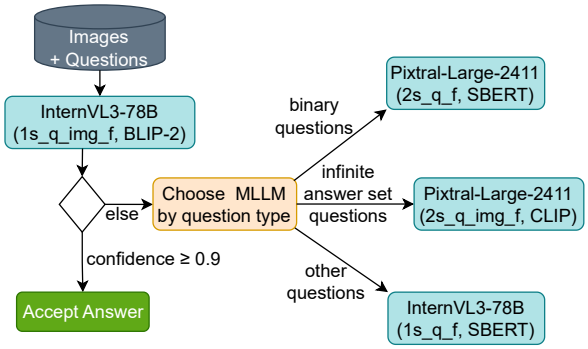

Figure 1: System overview. Abbreviations for few-shot example selection: #s = #-shot, q = question similarity, img = image similarity, f = filter for same figure type, nf = no filtering (search in entire train set).

2 Method

Our system is configurable to use different MLLMs in either a zero-shot or a few-shot setting. These settings are combined using an ensemble approach (see Figure 1) that first selects all high-confidence answers from a configuration that we find to be well-calibrated, i.e., the predicted confidence scores align well with the actual empirical accuracy on the development set. We approximate answer confidence by exponentiating the mean log-probability of all generated answer tokens. For the remaining instances, the model configuration is selected based on question type as identified on the development set. The MLLM is prompted with each image and the associated question (see Ap-

[1] https://github.com/coling-unia/few-shot-scivqa2025
[2] https://huggingface.co/datasets/katebor/SciVQA

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 230–239

Rank	Submission	R1-F1	RL-F1	BS-F1	Avg.
1.	ExpertNeurons	80.49	80.43	98.49	86.47
2.	THAii_LAB	78.99	78.92	98.39	85.43
3.	**Coling-UniA**	78.62	78.56	98.17	85.12
	Median	75.83	75.75	98.36	83.31

Table 1: Overview of SciVQA@SDP 2025 results. Metrics: R1 = ROUGE-1, RL = ROUGE-L, BS = BERTS.

pendix A.2). Following the oracle-style setup of the Shared Task, we also provide the model with additional image metadata that is included in the dataset, i.e., the image caption, figure type, and whether the image contains multiple subfigures. The task description depends on whether there are pre-defined answer options for the questions. The model is instructed to answer and to determine whether it is possible to answer based solely on the provided information.

To enhance the reproducibility, our selection of MLLMs is constrained to open-weights models. We use InternVL3-78B (Zhu et al., 2025) and Pixtral-Large-Instruct-2411.[3] We run all models using 16-bit quantization and a temperature of 0.

Few-shot Example Retrieval. We evaluate different few-shot retrieval approaches. First, we use question similarity to select examples from the training data for the input instance. For ranking, we use the cosine similarities of the questions' SBERT embeddings (Reimers and Gurevych, 2019). Second, we select examples based on the question-image similarity using CLIP (Radford et al., 2021). We compute CLIP embeddings for each question and image, normalize them, compute the mean embedding of each image-question pair, normalize again, and determine the best-fit example using cosine similarity. We also experimented with computing similarities based on the image-question embeddings directly provided by BLIP-2 (Li et al., 2023). In case of similarity ties, we choose the first instance in the order as they are provided in the training set.

For both settings, we retrieve few-shot examples from the training set in two variants: (1) We consider only the subset of the training data that has the same figure type, and, if possible, the same number of sub-figures, as the input instance. (2) We search for few-shot examples in the entire training set. In both cases, we exclude all instances that use the input image from the set of few-shot candidates. We do not filter training data based on the question type. In the oracle-style setting of the Shared Task, it would have been possible to additionally filter based on question type. We do so only indirectly by searching for questions and images with high embedding similarity, which makes our approach more directly applicable to real-world scenarios, where the question type may not be provided. Moreover, the question type "unanswerable" directly reveals the gold answer.

Our retrieval method ranks instances. It can thus be used to retrieve an arbitrary number of few-shot examples. We evaluate the performance of these retrieval strategies in one-shot and two-shot settings. When using two examples, the model is given one answerable and one unanswerable example.

3 Development Results and Ablations

Since our approach does not require any fine-tuning, we combine the training and validation sets into one development set. This section describes our careful experimentation and ablation studies on the development set. For more information on the dataset, see Appendix A.3.

We rely on the metrics of the Shared Task, F1, Precision, and Recall of ROUGE-1, ROUGE-L (Lin, 2004), and BERTScore (BERTS, Zhang et al., 2020), respectively, to evaluate our approach. However, we focus on ROUGE-1 F1, as the BERTS scores are similar for all approaches, and the ROUGE-L scores are comparable to ROUGE-1.

We run our experiments on Nvidia A100 (80 GB) GPUs, using up to 4 GPUs in parallel. The total amount of GPU hours was about 3600h.

3.1 Retrieval of Few-Shot Examples

As shown in Figure 2, the degree to which the question type of the retrieved few-shot examples matches that of the input instance varies greatly by question type. Searching for examples using only question similarity leads to matching the input instance's question type far more often than searching using image and question similarity. However, this does not seem to make a marked difference in overall performance. We found BLIP-2's text-image embeddings to primarily reflect the image content, resulting in many ties.[4]

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

[4] Our tie-breaking strategy leads to instances of the question type "closed-ended infinite answer set visual" to be selected, which comes first in the training set ordering for each image.

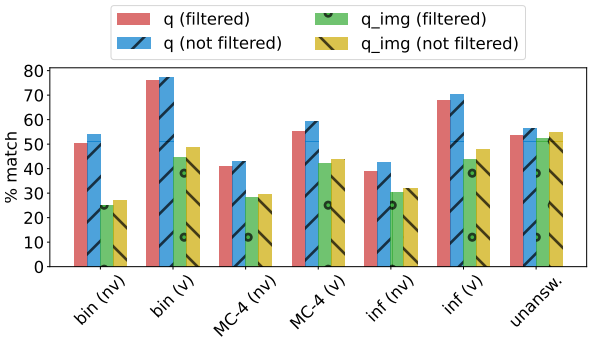

Figure 2: Percentage of selected one-shot example matching the question type of the input instance. bin = binary question, MC-4 = four answer options, inf = infinite answer set, unansw. = unanswerable, (v) = visual, (nv) = non-visual, filtered = filter for same figure type, not filtered = search in entire train set, q = question similarity, img = image similarity.

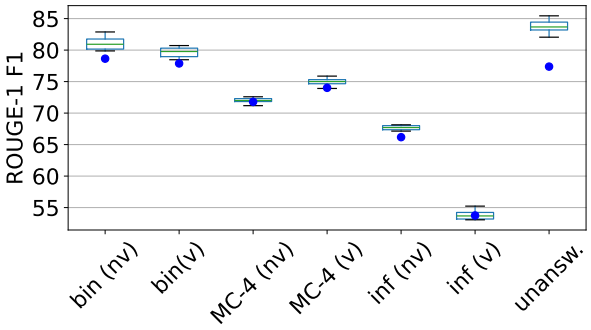

Figure 3: ROUGE-1 F1 scores per question type. Boxplot: 1-shot and 2-shot question and question+image similarity configurations of Pixtral-Large-2411. Blue dots = Pixtral-Large-2411 (0-shot).

3.2 Impact of Few-Shot Examples

Table 2 compares the effectiveness of InternVL3-78B and Pixtral-Large-2411 using various few-shot settings with that of our ensemble approaches. Adding few-shot examples generally improves performance. We cannot report 2-shot results for InternVL3-78B because its context window is too small to incorporate two examples. Comparing performance by question type reveals that adding examples can be highly beneficial, e.g., for recognizing unanswerable questions, though they can also be distracting (see Figure 3 or Appendix A.1). However, adding two examples is almost always beneficial. Furthermore, using one answerable and one unanswerable example helps the model to distinguish between these two types of instances, especially when compared to using only one example (see Table 3).

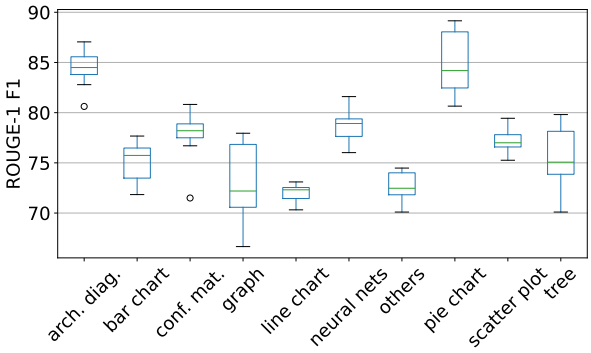

Figure 4: ROUGE-1 F1 scores per figure type of all configurations (zero- and few-shot) of both MLLMs visualized as boxplots.

3.3 Question/Figure Type Ensemble

To determine the best configuration of MLLM and few-shot strategy for each pair of question type and figure type, we systematically search for the optimal ensemble settings by obtaining and analyzing distributions of performance scores over subsets of the data similar to cross-validation.

While there appears to be a general trend of enhanced performance with the use of examples (see Table 2), our findings reveal considerable variations in the performance of our configurations across different question and figure types (see Figure 3 and Figure 4 or Appendix A.1). Therefore, we use the results on the development set to systematically identify the optimal combination of configurations that work well across as many subsets of the data as possible. The dataset consists of seven evenly represented question types and various figure types that are not evenly distributed. We record performance scores for each figure type separately. To avoid overfitting, we summarize all figure types that encompass less than two percent of the total number of figures into the figure type "others", which leads to nine groups with homogeneous figure types (line chart, tree, scatter plot, pie chart, bar chart, architecture diagram, neural networks, confusion matrix, graph) plus one group of the "others", i.e., 10 groups in total. For the largest figure type, i.e., line chart, we divide the data into seven groups by further dividing the data by question type. In total, we divide the data into 16 groups (8 homogeneous figure types, 1 "others", and 7 subsets with line charts).

We split the data of each group into 5 folds and compute performance scores. We repeat this process at least 10 times with different splits until the predicted best-performing configuration remains

Setting	Configuration	R1-F1	R1-P	R1-R	RL-F1	RL-P	RL-R	BS-F1	BS-P	BS-R
Individual runs (dev set)	InternVL (0s)	74.2	75.2	74.9	74.1	75.1	74.8	97.1	97.3	97.0
	InternVL (1s_q_f)	74.7	75.7	74.8	74.6	75.6	74.8	97.8	97.9	97.8
	InternVL (1s_q_nf)	74.5	75.5	74.6	74.4	75.4	74.6	97.8	97.8	97.7
	InternVL (1s_q_img_f)	74.8	75.6	**75.2**	74.7	75.6	**75.1**	97.8	97.8	97.8
	InternVL (1s_q_img_nf)	74.7	75.7	75.1	74.6	75.6	75.0	97.8	97.8	97.8
	InternVL (1s_q_img_f, BLIP2)	**75.0**	**76.0**	**75.2**	**74.9**	**76.0**	**75.1**	**97.9**	**98.0**	**97.9**
	Pixtral (0s)	71.4	72.5	72.4	71.2	72.4	72.2	96.3	96.6	96.0
	Pixtral (1s_q_f)	72.8	74.0	73.1	72.7	73.9	73.0	97.5	97.6	97.5
	Pixtral (1s_q_nf)	72.3	73.5	72.6	72.2	73.4	72.5	97.5	97.5	97.5
	Pixtral (1s_q_img_f)	72.8	74.1	73.1	72.7	74.0	73.0	97.4	97.5	97.3
	Pixtral (1s_q_img_nf)	72.8	74.0	73.2	72.7	73.9	73.1	97.4	97.5	97.3
	Pixtral (2s_q_f)	73.9	75.2	74.0	73.8	75.1	73.9	97.7	97.8	97.7
	Pixtral (2s_q_nf)	73.7	75.0	73.8	73.6	74.9	73.7	97.7	97.8	97.7
	Pixtral (2s_q_img_f)	74.1	75.5	74.2	74.0	75.4	74.1	97.7	97.8	97.6
	Pixtral (2s_q_img_nf)	73.8	75.2	73.9	73.7	75.1	73.8	97.6	97.8	97.6
Ensembles (dev set)	Question/Figure-Type Ensemble	76.6	78.0	76.5	76.5	77.8	76.4	97.9	98.0	**97.9**
	Confidence-Informed Ensemble	**76.9**	**78.2**	**76.8**	**76.8**	**78.1**	**76.8**	**98.0**	**98.1**	**97.9**
Results on test set	InvernVL (1s_q_img_f, BLIP2)	77.2	78.0	77.4	77.2	77.9	77.3	98.1	98.2	**98.1**
	Question/Figure-Type Ensemble	77.7	78.8	77.7	77.6	78.7	77.6	98.1	98.2	98.0
	Confidence-Informed Ensemble	**78.6**	**79.7**	**78.6**	**78.6**	**79.6**	**78.5**	**98.2**	**98.3**	**98.1**

Table 2: Results of individual runs vs. ensembles on development and test set. Abbreviations for few-shot example selection: #s = #-shot, q = question similarity, img = image similarity, f = filter for same figure type, nf = no filtering (search in entire train set). Metrics: R1 = ROUGE-1, RL = ROUGE-L, BS = BERTS, P = Precision, R = Recall. Question/Figure-Type Ensemble refers to the approach described in section 3.3 and Confidence-Informed Ensemble to that of section 3.4.

Approach	Precision
Pixtral (0s)	93.0
Pixtral (1s_q_f)	89.2
Pixtral (1s_q_img_f)	89.3
Pixtral (1s_q_img_nf)	90.3
Pixtral (1s_q_nf)	88.7
Pixtral (2s_q_f)	92.7
Pixtral (2s_q_img_f)	94.1
Pixtral (2s_q_img_nf)	93.7
Pixtral (2s_q_nf)	93.3

Table 3: Precision of instances predicted to be unanswerable.

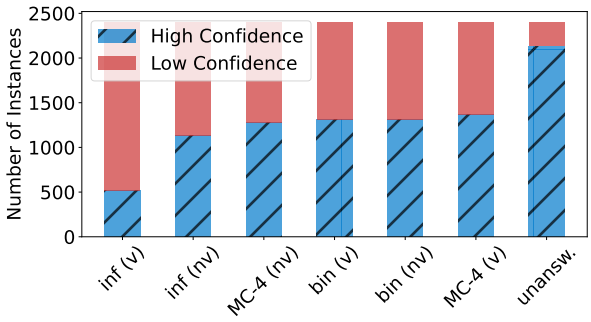

Figure 5: Number of instances having received high confidence answer of InternVL3-78B (1s_q_img_f, BLIP) by question type.

constant. In each fold, we calculate the ROUGE-1 F1 score for all configurations, then subtract the highest score achieved in that fold. For each configuration, we then compute the mean of these scores across all folds and all runs. The best-performing configuration is identified by the highest score. For the final chosen configuration of this ensemble, refer to Table 8 in Appendix A.1.

3.4 Confidence-Informed Ensemble

Figure 6 shows that InternVL3-78B (1s_q_img_f, BLIP2) with examples derived from BLIP-2, which focus primarily on image similarity as explained in Sec. 3.1, is meaningfully calibrated. This means that high confidence scores indicate highly likely correct instances (refer to Appendix A.1 for detailed results). Thus, for our final submission, we directly use all predictions from this model with a confidence score of at least 90%, which corresponds to approximately half of the instances, in the initial stage. As shown in Figure 5, the number of high-confidence instances varies by question type. The model is most confident on identifying unanswerable questions, while it is least sure about its answers for questions with infinite answers sets about the image's visual features. After removing high-confidence instances, the performance per question type varies widely between our configruations (see Appendix A.1 for detailed results). The best configuration per question type does not seem

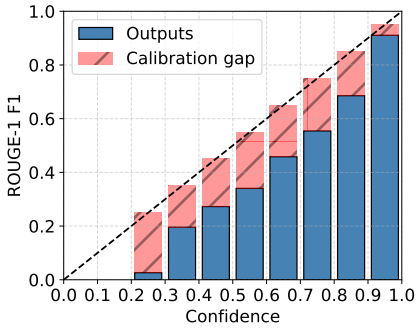

Figure 6: Calibration plot for InternVL3-78B (1s_q_img_f, BLIP) showing that instances with confidence score ≥ 0.9 have high expected accuracy.

to depend on whether the question incorporates visual or non-visual features. Since the vast majority of the remaining instances are of figure type *line chart*, we do not perform cross-validation to determine the optimal configuration of approaches. Instead, we select the best-performing approach for each question type, while also trying to reduce the number of approaches required.

As can be seen in Figure 1, we use the following models for the remaining instances: Pixtral-Large-2411 (2s_q_f) for binary questions, Pixtral-Large-2411 (2s_q_img_f) for questions with an infinite answer set, and InternVL3-78B (1s_q_f) for all others.

4 Results on Test Set

Table 2 also shows the results of our approaches on the test set, indicating that our ensembling strategies improve the performance compared to using only one approach to answer all questions. On the test set, we also find the confidence-informed ensemble to work best, while the question/figure type ensemble outperforms the simple InternVL model not as strongly as on the development set. The confidence-informed ensemble is the approach submitted for the leaderboard, ranking third in the official evaluation (almost on par with the second-ranking system) as shown in Table 1, and outperforming the baseline by about 4 percentage points.

5 Discussion and Conclusion

This paper described our submission to the SciVQA 2025 Shared Task. Our results show that MLLMs are highly effective at answering questions about scientific figures. However, performance varies greatly by question type. Results on finite answer sets are considerably better than on infinite ones. In particular, answering infinite answer set questions about visual features of images remains challenging, highlighting the need for a more sophisticated approach.

The use of few-shot examples improves performance. However, there are no major performance differences between retrieving the examples by question or question-image similarity.

Limitations

Since the ACL-Fig and SciGraphQA datasets, on which the figures in this Shared Task are based, rely on images published several years ago, some of these images may have already been exposed to MLLMs during training.

Another limitation is performance on unanswerable questions. Although our approach performed best on this question type, it is difficult to determine if it would perform equally well on real-world unanswerable questions. This is because the unanswerable questions in this dataset follow a different pattern than the answerable ones. For example, they mostly refer to material unavailable to the model and often do not focus on the images' visual/non-visual features.

Acknowledgments

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the BayernKI project v110ee. BayernKI funding is provided by Bavarian state authorities.

References

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. 2025. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

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. *CoRR*, abs/2311.16483.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. Acl-fig: A dataset for scientific figure classification. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 37th AAAI Conference on Artificial Inteligence (AAAI 2023), Remote, February 14, 2023*, volume 3656 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *CoRR*, abs/2308.03349.

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, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. 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.

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.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1516–1525. IEEE.

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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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*. Association for Computational Linguistics.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *CoRR*, abs/2402.12185.

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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1882–1898. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Detailed Results on Development Set

Table 4 shows the detailed results of the different zero- and few-shot approaches on the development set, broken down by question type. Performance varies greatly between question types, indicating that questions with an infinite answer set are more difficult. Furthermore, performance depends on the MLLM and few-shot configuration used. Mostly, using examples is beneficial for performance.

As shown in Table 5, the performance of the different configurations also depends on the figure type of the image.

Table 6 reports the ROUGE-1 F1 score per confidence bin and the relative proportion of respective bin of the development set. Interestingly, the configuration that uses BLIP-2 to retrieve similar examples is well-calibrated for high confidence. In

general, InternVL3-78B appears to be better calibrated than Pixtral-Large-2411 for our task.

The performance of different approaches per question type can be seen in Table 7 after having removed all instances of InternvL3-78B (1s_q_img_f, BLIP) with a confidence of $\geq 90\%$. Performance is worse compared to Table 4 since the high confidence answers are removed. Nevertheless, there are still large performance differences between the different approaches.

Table 8 shows the best configurations per figure and question type identified via cross-validation for the Question/Figure Type Ensemble.

A.2 Detailed Prompt

Figure 7 shows the prompt used in our approach. Its formatting depends on the annotated metadata, i.e., whether the instance has annotated answer options and whether the figure consists of multiple subfigures.

A.3 Dataset Characteristics

The dataset consists of 3000 real-world figures extracted from English scientific publications available in the ACL Anthology and arXiv. The figures can be categorized into different figure types such as *line chart*, *tree*, or *scatter plot*. These figure types are not evenly distributed. For example, *line chart* makes up 65% of all figures in the development set (see Figure 8).

Each figure is annotated with seven questions. Two binary questions (one focusing on visual features and one focusing on non-visual features), two questions with four answer options respectively (one visual and one non-visual), two questions with infinite answer sets (one visual and one non-visual), and one unanswerable question. The unanswerable questions are not subdivided into visual and non-visual questions, and they generally follow a different pattern than the answerable ones. For example, they mostly refer to material unavailable to the model and often do not focus on the images' visual/non-visual features.

Approach	binary (nv)	binary(v)	MC-4 (nv)	MC-4 (v)	inf (nv)	inf (v)	unanswerable
InternVL (0s)	78.9	80.9	76.6	79.3	66.4	51.2	86.2
InternVL (1s_q_f)	82.0	80.8	**76.9**	**79.9**	63.4	49.3	90.4
InternVL (1s_q_nf)	82.0	81.0	76.3	79.1	63.5	50.0	89.5
InternVL (1s_q_img_f)	80.4	**81.3**	76.3	79.3	65.3	49.5	**91.2**
InternVL (1s_q_img_nf)	80.6	80.8	76.3	78.6	65.7	50.1	91.0
InternVL (1s_q_img_f, BLIP)	82.7	80.6	76.7	79.0	67.4	51.8	87.0
Pixtral (0s)	78.6	77.9	71.8	74.0	66.2	53.8	77.4
Pixtral (1s_q_f)	80.2	79.5	71.7	73.9	67.4	53.4	83.4
Pixtral (1s_q_nf)	79.9	78.5	71.2	74.1	67.5	53.1	82.0
Pixtral (1s_q_img_f)	80.3	79.0	71.9	75.0	67.1	53.1	83.2
Pixtral (1s_q_img_nf)	79.9	79.0	72.3	75.1	67.3	53.2	83.0
Pixtral (2s_q_f)	**82.9**	80.7	72.1	74.9	**68.1**	54.0	84.4
Pixtral (2s_q_nf)	82.1	80.3	71.9	75.0	68.0	54.6	84.0
Pixtral (2s_q_img_f)	81.5	80.4	72.6	75.8	67.9	**55.2**	85.4
Pixtral (2s_q_img_nf)	81.6	80.0	72.4	75.9	**68.1**	54.1	84.5

Table 4: Results (ROUGE-1 F1 scores) on development set by question type. v=visual, nv= non-visual.

Approach	architecture diagram	bar chart	confusion matrix	graph	line chart	neural networks	others	pie chart	scatter plot	tree
InternVL (0s)	83.9	**77.7**	77.0	75.4	72.2	78.9	73.4	87.4	78.0	76.2
InternVL (1s_q_f)	86.2	77.3	76.7	76.8	72.5	79.2	**74.5**	**89.2**	75.6	78.6
InternVL (1s_q_nf)	85.0	76.4	78.6	77.4	72.5	79.2	73.4	88.0	76.2	77.8
InternVL (1s_q_img_f)	86.2	76.6	78.4	77.7	72.6	78.8	74.4	88.7	77.1	78.2
InternVL (1s_q_img_nf)	85.4	76.1	78.3	**78.0**	72.6	79.7	73.9	88.0	76.9	**79.2**
InternVL (1s_q_img_f, BLIP)	**87.0**	76.3	78.8	77.1	**72.8**	**81.6**	74.4	88.1	77.9	78.1
Pixtral (0s)	80.6	72.8	71.5	66.7	70.3	76.0	70.1	82.8	75.3	70.1
Pixtral (1s_q_f)	84.7	73.8	77.5	67.9	71.3	77.5	72.0	82.1	76.7	73.5
Pixtral (1s_q_nf)	83.3	73.4	77.7	69.8	71.0	77.1	70.3	81.6	75.7	72.6
Pixtral (1s_q_img_f)	82.8	73.3	78.1	70.7	71.4	77.7	71.4	80.9	76.9	73.9
Pixtral (1s_q_img_nf)	84.1	71.8	78.0	70.3	71.5	76.2	71.6	80.6	76.9	74.3
Pixtral (2s_q_f)	83.6	76.2	79.9	71.4	72.3	79.6	72.7	82.6	77.4	74.5
Pixtral (2s_q_nf)	83.9	73.5	77.4	72.6	72.3	78.6	72.3	83.7	77.8	74.2
Pixtral (2s_q_img_f)	85.2	75.4	80.4	71.7	72.5	78.9	71.9	84.4	**78.9**	75.7
Pixtral (2s_q_img_nf)	84.3	74.2	**80.8**	71.3	72.3	79.3	72.2	84.0	77.6	73.8

Table 5: Results (ROUGE-1 F1 scores) on development set by figure type.

Approach	0.3_0.4	0.4_0.5	0.5_0.6	0.6_0.7	0.7_0.8	0.8_0.9	0.9_1.0
InternVL (0s)	15.2 (0.1)	**34.5** (1.2)	**48.7** (5.3)	**55.4** (13.6)	**69.9** (21.2)	**73.3** (18.2)	87.9 (40.4)
InternVL (1s_q_f)	18.5 (0.1)	30.2 (0.8)	30.3 (2.9)	45.4 (7.4)	56.8 (14.1)	66.4 (16.2)	88.0 (58.5)
InternVL (1s_q_nf)	22.9 (0.1)	25.0 (0.9)	31.1 (3.1)	45.5 (7.2)	55.5 (14.1)	66.3 (16.2)	88.1 (58.4)
InternVL (1s_q_img_f)	21.2 (0.1)	30.9 (0.8)	37.6 (3.4)	48.8 (8.4)	57.6 (14.8)	68.6 (16.5)	88.1 (55.9)
InternVL (1s_q_img_nf)	15.9 (0.1)	28.3 (0.8)	38.7 (3.3)	48.1 (8.4)	56.1 (14.7)	69.2 (16.7)	88.3 (55.9)
InternVL (1s_q_img_f, BLIP)	19.6 (0.2)	27.3 (1.0)	34.0 (3.5)	45.8 (8.3)	55.4 (15.3)	68.6 (17.8)	**91.1** (53.9)
Pixtral (0s)	**25.0** (0.0)	24.7 (0.3)	34.6 (1.8)	45.4 (6.0)	61.2 (17.3)	64.5 (24.6)	83.0 (50.0)
Pixtral (1s_q_f)	0.0 (0.0)	31.1 (0.3)	33.7 (1.6)	43.3 (5.6)	54.6 (13.4)	61.8 (19.9)	84.7 (59.2)
Pixtral (1s_q_nf)	11.4 (0.0)	32.6 (0.3)	32.8 (1.6)	42.1 (5.6)	52.9 (13.5)	61.7 (19.9)	84.5 (59.1)
Pixtral (1s_q_img_f)	16.3 (0.0)	25.8 (0.3)	39.5 (1.8)	45.5 (6.2)	55.9 (13.8)	62.9 (20.8)	84.8 (57.1)
Pixtral (1s_q_img_nf)	19.0 (0.0)	23.5 (0.3)	40.0 (1.7)	46.9 (6.2)	54.3 (14.0)	63.2 (20.5)	85.0 (57.1)
Pixtral (2s_q_f)	0.0 (0.0)	24.1 (0.2)	37.8 (1.2)	40.8 (4.6)	52.8 (12.3)	62.1 (19.2)	84.9 (62.5)
Pixtral (2s_q_nf)	0.0 (0.0)	20.1 (0.2)	31.4 (1.2)	42.3 (4.6)	52.3 (11.8)	60.9 (19.6)	85.0 (62.6)
Pixtral (2s_q_img_f)	14.3 (0.0)	25.6 (0.1)	33.6 (1.4)	47.2 (5.2)	55.1 (12.6)	61.9 (20.3)	85.6 (60.4)
Pixtral (2s_q_img_nf)	0.0 (0.0)	28.8 (0.2)	34.2 (1.4)	42.8 (5.1)	52.6 (12.8)	63.1 (19.9)	85.4 (60.7)

Table 6: Results (ROUGE-1 F1 scores) per confidence bin. The values in brackets indicate the relative proportion of instances in each bin.

Approach	binary (nv)	binary(v)	MC-4 (nv)	MC-4 (v)	inf (nv)	inf (v)	unanswerable
InternVL (0s)	64.8	68.8	60.6	62.0	50.6	45.0	34.3
InternVL (1s_q_f)	70.7	68.3	**61.0**	**63.1**	46.9	42.0	48.6
InternVL (1s_q_img_f)	67.8	69.2	60.1	62.2	49.9	42.3	50.2
InternVL (1s_q_img_nf)	68.2	67.9	59.6	60.8	50.1	43.0	**50.8**
InternVL (1s_q_nf)	70.2	68.5	60.4	62.0	47.3	42.8	42.8
Pixtral (0s)	67.8	67.1	58.1	58.4	52.3	48.1	29.8
Pixtral (1s_q_f)	70.3	68.9	57.2	57.7	53.8	48.0	38.2
Pixtral (1s_q_img_f)	71.4	69.0	57.6	59.2	53.8	48.0	32.2
Pixtral (1s_q_img_nf)	70.3	68.0	58.8	59.3	53.8	48.1	33.1
Pixtral (1s_q_nf)	70.1	67.0	56.8	58.1	53.6	47.8	34.7
Pixtral (2s_q_f)	**74.0**	**70.8**	57.5	58.9	54.8	48.4	36.8
Pixtral (2s_q_img_f)	72.5	69.4	58.2	59.9	55.2	**49.5**	39.0
Pixtral (2s_q_img_nf)	72.7	69.0	58.4	60.0	**55.4**	48.4	35.8
Pixtral (2s_q_nf)	73.2	69.9	57.4	59.0	54.5	**49.5**	37.0

Table 7: Results (ROUGE-1 F1 scores) on development set by question type after removing high confidence instances of run with InternVL3-78B (1s_q_img_f with BLIP-2).

Figure Type	inf (v)	inf (nv)	bin (v)	bin (nv)	MC-4 (v)	MC-4 (nv)	unansw.
line chart	Pixtral (2s_q_img_f)	Pixtral (2s_q_nf)	Pixtral (2s_q_f)	Pixtral (2s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_img_f)
tree	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)
scatter plot	Pixtral (2s_q_img_f)	Pixtral (2s_q_img_f)	Pixtral (2s_q_img_f)	Pixtral (2s_q_img_f)	Pixtral (2s_q_img_f)	Pixtral (2s_q_img_f)	Pixtral (2s_q_img_f)
pie chart	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)
bar chart	InternVL (0s)	InternVL (0s)	InternVL (0s)	InternVL (0s)	InternVL (0s)	InternVL (0s)	InternVL (0s)
architecture diagram	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)
neural networks	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)
confusion matrix	Pixtral (2s_q_img_nf)	Pixtral (2s_q_img_nf)	Pixtral (2s_q_img_nf)	Pixtral (2s_q_img_nf)	Pixtral (2s_q_img_nf)	Pixtral (2s_q_img_nf)	Pixtral (2s_q_img_nf)
graph	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)	InternVL (1s_q_img_nf)
others	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)	InternVL (1s_q_f)

Table 8: Best configurations for combination of figure type and question type identified via cross-validation for Question/Figure Type Ensemble.

System Message: You are an assistant answering questions about (semi-)structured figures such as charts and diagrams. Answer the question as precisely as possible.
User Message: Image: {image}
Question: '{question}'

```
if image_metadata['answer_options']:
    Answer options: {answer_options}
```

Additional Information:

- The caption of the image is 'image_metadata['caption']'.

```
if image_metadata["compound"]:
    - The figure image contains {image_metadata['figs_numb']} (sub)figures which can be separated and constitute individual figures.
else:
    - The figure image contains a single figure object which cannot be decomposed into multiple subfigures.
```

- The figure type is '{image_metadata['figure_type']}'.

Task:
You are presented with a figure and an associated question.

```
if image_metadata['answer_options:']:
    Your task is to select the correct answer options based on the figure. One or more answer options are correct. Only respond with
    the key(s) of the correct answer option(s), so e.g., 'A,C' if answer options A and C are correct.
else:
    Your task is to answer the question based on the figure.
```

You should only use the information in the figure to answer the question. Do not use any external knowledge or information. If the figure does not provide enough information to answer the question, respond with 'It is not possible to answer this question based only on the provided data.'. If you can answer the question, simply provide the answer without further explanation and do not repeat the question.
Answer:

Figure 7: The zero-shot prompt is formatted based on the annotated metadata via conditional statements. The MLLM is not given the if-else logic; it is only given the indented text inside the block. Bold text is not part of the prompt for the LLM either; it only indicates which parts of the prompt belong to the system or user message. Values in brackets are placeholders for the respective instance's actual values.

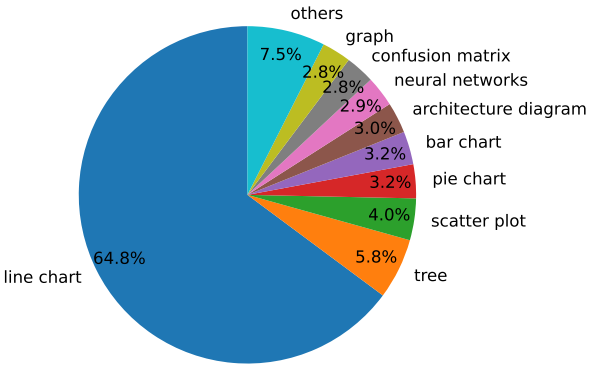

Figure 8: Figure type distribution on development set.