Encoded but Not Routed: Explaining the Table-Chart Gap in Scientific Claim Verification

Sunisth Kumar[1,2] **Xanh Ho**[2] **Tim Schopf**[3]
Andre Greiner-Petter[3,4] **Florian Boudin**[5] **Akiko Aizawa**[1,2,3]

[1]The University of Tokyo [2]NII LLMC [3]National Institute of Informatics
[4]University of Göttingen [5]Inria, LS2N, Nantes Université

sunisth@g.ecc.u-tokyo.ac.jp {xanh,aizawa}@nii.ac.jp tim.schopf@t-online.de
greinerpetter@gipplab.org florian.boudin@univ-nantes.fr

Abstract

Multimodal LLMs are increasingly used to assist scientific peer review, where a core requirement is verifying whether claims in a paper are supported by its evidence. Prior work has shown that models perform substantially better at this task when the evidence is a table than when it is a chart of the same underlying data. This raises the question of whether models fail to extract information from charts, or do they extract it but fail to use it when forming their prediction? We study this question through layer-wise linear probing and attention analysis on three open-weight VLMs over table and chart evidence, representing the same underlying data. We find consistent evidence for the latter. Chart information is encoded in the models' intermediate representations but does not reach the prediction position, a gap that is absent for tables and holds across all conditions tested. Attention analysis further reveals that this disconnect takes two architecturally distinct forms across model families. These findings point toward reframing the table-chart gap as a failure of how encoded visual information is used at prediction time, rather than a failure of encoding itself. [1]

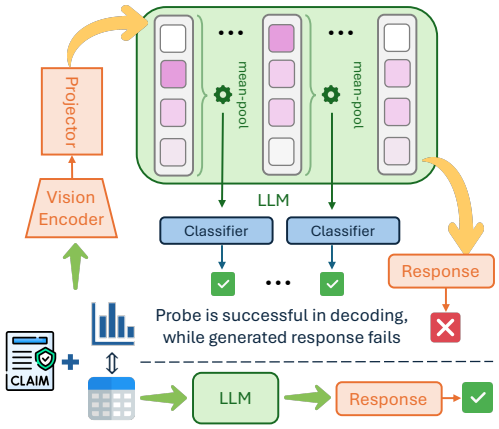


Figure 1: Overview of our approach. We apply linear probing to compare how chart and table evidence is encoded and passes to the model response.

1 Introduction

Multimodal large language models are increasingly used for scientific claim verification, a task that requires determining whether a given claim is supported or refuted by evidence from research papers (Lu et al., 2023; Wang et al., 2025; Ho et al., 2026b). In practice, the same experimental finding may be presented as a table in one case and as a bar or line chart in another. A robust verifier should therefore make consistent decisions across evidence formats when the underlying data is unchanged.

Recent work shows that current multimodal LLMs perform substantially better on tables than on equivalent charts (Ho et al., 2026a). Since the claim and the underlying data are held constant across formats, this reveals a fundamental brittleness in how models process semantically equivalent information in different modalities. Understanding why this gap exists matters because it determines whether models fail to extract visual information or fail to use information they have already extracted.

What explains this asymmetry? One possibility is that models fail to perceive chart content, such that visual information does not reach model representations in a usable form. Another is that chart content is encoded in model representations but is not effectively routed to the prediction stage. These two explanations call for different responses. Perceptual failures point toward better vision encoders or chart-specific pre-training (Masry et al., 2022; Liu et al., 2025a), while utilization failures call for interventions on how encoded representations are used at prediction time (Zhang et al., 2025; Kawasaki et al., 2026). Distinguishing between them requires looking inside the model.

We take a diagnostic approach to this ques-

[1]https://github.com/ksunisth/encoded-not-routed

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3700–3709
November 6-10, 2026 ©2026 Association for Computational Linguistics

tion, studying three open-weight multimodal LLMs from two model families on a content-controlled SciTabAlign+ benchmark (Ho et al., 2026a) in which each claim is paired with semantically equivalent table and chart evidence. We analyze internal representations through linear probing (Belinkov, 2022) to understand whether chart failures arise from missing internal signals or from failures to use that signal during prediction (see Figure 1).

Our results reveal that chart-relevant signal is recoverable from intermediate representations yet is not effectively routed to final verification decisions, indicating a routing failure rather than a perceptual one.

2 Related Work

Scientific Claim Verification. Claim verification has been studied from textual evidence (Thorne et al., 2018), and extended to the scientific domain through SciFact (Wadden et al., 2020), SciTab (Lu et al., 2023), and SciTabAlign (Ho et al., 2025). Recent works have broadened the task to multimodal settings spanning tables and figures (Lal et al., 2025; Wang et al., 2025; Ho et al., 2026b), where multimodal LLMs have been shown to perform substantially better on tables than on equivalent charts (Ho et al., 2026a). Building on this finding, we investigate why that gap arises.

Chart Understanding and Interpretability in Multimodal LLMs. ChartQA (Masry et al., 2022) and CharXiv (Wang et al., 2024) benchmark VLM performance on chart-specific tasks, both revealing persistent gaps between VLM and human performance. (Dong and Crisan, 2026) further apply activation maps to visualize which chart regions models attend to. Unlike these studies, we hold the underlying data constant across formats, isolating format sensitivity from content variation. Recent work shows VLM failures may arise even when relevant visual information is present in representations (Kawasaki et al., 2026), and that models may attend to relevant visual evidence while still answering incorrectly (Liu et al., 2025b; Esmaeilkhani and Latecki, 2026). We study this representation-response gap in a format-controlled claim verification setting, where the same underlying data appears as table or chart evidence.

3 Experimental Setup

Task and Data. We study binary scientific claim verification on SciTabAlign+. The dataset pairs 162 unique scientific claims with a structured table and four chart variants (basic bar chart, symbol bar chart, line chart, and swapped chart), all drawn from the same underlying data. In the dataset, tables are provided as JSON and charts as JPG images. Each claim is labeled as *Supported* or *Refuted*, resulting in 372 table-based and 648 chart-based instances. Holding evidence content constant across formats allows us to attribute behavioral differences to format rather than to content.

Models. We study three open-weight multimodal LLMs spanning two families. We use Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-32B-Instruct from the Qwen family (Bai et al., 2025), and InternVL3-8B from the InternVL family (Zhu et al., 2025).

Linear Probing. We train a linear probe at every layer l of each model using leave-one-out cross-validation over the 162 claims, minimizing binary cross-entropy with Adam (Kingma and Ba, 2015) (learning rate 0.01, 300 steps). Each probe maps a hidden state $\mathbf{h}^{(l)}$ to a label prediction:

$$p\Big(y = 1 \mid \mathbf{h}^{(l)}\Big) = \sigma\Big(\mathbf{w}^\top \mathbf{h}^{(l)} + b\Big) \qquad (1)$$

Within each fold, we determine the probe's direction from a stratified 20% inner split of the training data. We evaluate each probe under two settings. In the *last-token* setting, $\mathbf{h}^{(l)}$ is the hidden state at the final input token position, from which the model generates its prediction. In the *mean-pool* setting, $\mathbf{h}^{(l)}$ is the mean hidden state averaged uniformly over all input token positions, which captures task-relevant information distributed across layers rather than concentrated at the final position (Skean et al., 2025). As a control for claim-text leakage, we train identical probes on claim text alone, passing each claim through the model without evidence input. See Appendix E.

Attention Analysis. At each layer l, we compute the fraction of last-token attention directed to image token positions, $a^{(l)}_{\text{img}}$, normalized by the proportional baseline:

$$\hat{a}^{(l)} = \frac{a^{(l)}_{\text{img}}}{|T_{\text{img}}| \,/\, |T|} \qquad (2)$$

where $|T_{\text{img}}|$ is the number of image tokens and $|T|$ is the total token count. We compute $a^{(l)}_{\text{img}}$ by averaging attention uniformly across heads and summing across all image-token positions, which measures total probability mass directed to the image

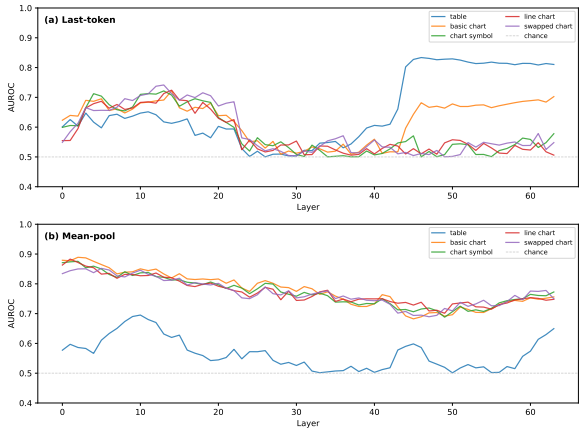


Figure 2: Probe AUROC by layer for Qwen2.5-VL-32B on SciTabAlign+. **(a)** Last-token: table rises sharply in late layers, chart variants peak early and fall toward chance in late layers. **(b)** Mean-pool: chart signal remains decodable across all layers, but does not reach prediction position. Results for all models are shown in Appendix B.2.

region. A value of $\hat{a}^{(l)} = 1.0$ means the model attends to image tokens in proportion to their count in the input. Values below 1.0 indicate less attention to image content than this baseline.

Chain-of-Thought Ablation. We test whether explicit chart verbalization before prediction can improve performance accuracy by prompting each model to describe the chart before predicting the label (see Appendix D).

Evaluation. We evaluate probing using area under the receiver operating characteristic curve (AUROC). Our interest is in whether task-relevant information is decodable from representations, and AUROC captures this without relying on a fixed classification threshold. We also report macro-F1 and accuracy for direct comparison with model output accuracy.

4 Results

Chart signal is recoverable from intermediate representations. Figure 2 shows that mean-pooled probes substantially outperform last-token probes on chart variants across all models. For Qwen2.5-VL-32B, mean-pool AUROC on basic chart reaches 88.9% compared to 72.2% at the last token, indicating that chart-relevant signal is distributed across hidden states but has not concentrated at the prediction position. Notably, mean-pool AUROC is higher for chart variants than for table evidence across all models (84–89% vs. 65–70%), the reverse of the last token pattern in Table 1. This inversion suggests that chart information is more broadly distributed across hidden states than table information, but specifically fails to concentrate at the prediction position where table information already aggregates. Claim-only probes trained on claim text without evidence peak at 58–67% AUROC across models (see Appendix E), well below the mean-pool AUROC on chart variants, ruling out claim-text leakage and indicating that the probe captures information present in image tokens but absent from claim text alone. Table and chart also converge to high cosine similarity in deeper layers (see Appendix C), ruling out representational divergence as the source of the behavioral gap.

Encoded signal does not reach the prediction position. Table 1 shows that last-token AUROC is overall consistently higher for table evidence than for chart evidence across all models and chart variants. As shown in Figure 3, a linear classifier trained on mean-pooled representations exceeds model inference accuracy on all chart variants across all three models, a pattern absent for table evidence. The dissociation is significant (McNemar's test (McNemar, 1947), $p < 0.05$ in all conditions and $p < 0.01$ in all but one) and reverses on table evidence, confirming format specificity. For InternVL3-8B, the mean-pool probe achieves an accuracy of 82.7% on basic chart while the model itself reaches 56.2%. This dissociation, where a simple linear classifier outperforms the model using its own representations, suggests that chart-relevant information is present but is not effectively leveraged at prediction time. The failure is therefore not one of encoding but of how encoded information is used when forming a prediction.

	Qwen2.5-VL-7B		Qwen2.5-VL-32B		InternVL3-8B	
Format	AUROC	F1	AUROC	F1	AUROC	F1
Table	**72.7**	**66.1**	**83.4**	**75.2**	**72.2**	**64.8**
Basic chart	61.9	59.0	72.2	67.6	63.3	58.5
Chart symbol	62.0	60.3	72.2	70.2	58.8	55.5
Line chart	57.2	56.1	72.4	67.7	61.8	62.8
Swapped chart	60.8	59.1	74.2	67.4	56.8	53.7

Table 1: Best last-token probe AUROC and macro-F1 across all layers on SciTabAlign+.

Failure patterns differ between Qwen and InternVL. Figure 4 shows that chart underutilization takes different forms across the two model families. Qwen-family models attend to image tokens only 4–11% of the proportional baseline at

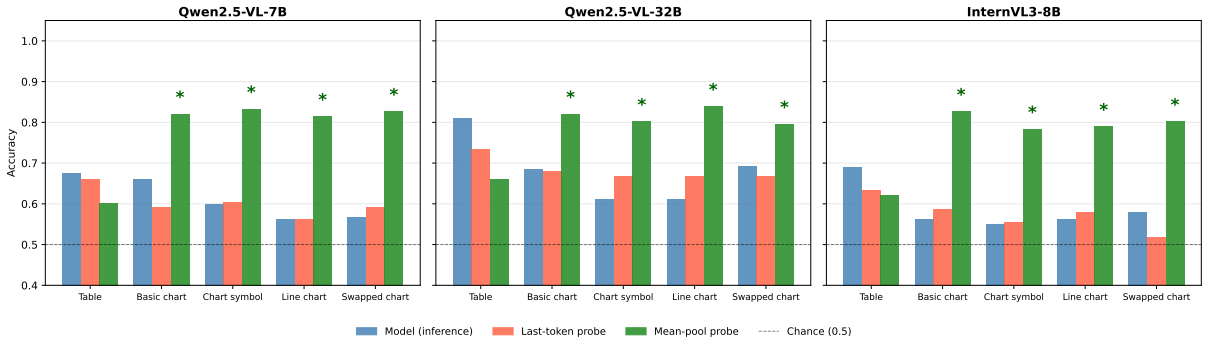


Figure 3: Probe accuracy vs. model inference accuracy on SciTabAlign+. The $*$ indicates chart variants where the mean-pool probe exceeds model accuracy.

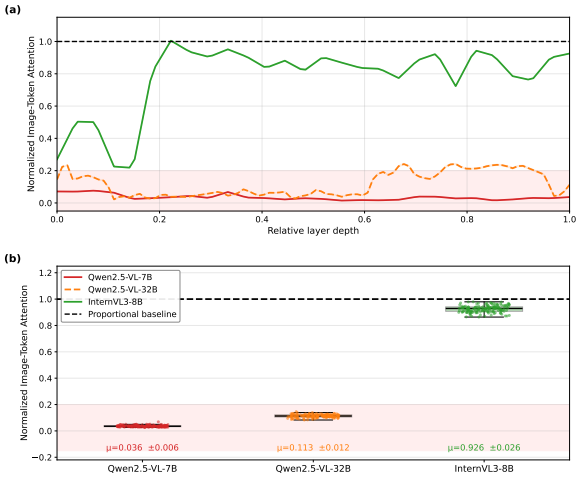


Figure 4: Image-token attention relative to the proportional baseline across layers on SciTabAlign+ (*basic chart* variant). The shaded region indicates attention below 20% of the proportional baseline. **(a)** Attention trajectories across layers. **(b)** Distribution of final-layer relative attention per claim.

the final layer, suggesting an *attention routing failure* in which chart evidence is largely bypassed when predicting the claim label. InternVL3-8B, by contrast, maintains near-proportional image-token attention throughout most of the layers (93% of the proportional baseline), yet still fails to outperform its own mean-pool probe on chart variants, consistent with a *post-attention integration failure*, where the model attends to chart evidence but does not integrate it into the prediction. As shown in Appendix B.3, image-token attention distributes across positions rather than concentrating in summary tokens, confirming distributed visual routing. As shown in Table 4, chain-of-thought prompting further distinguishes the two families. CoT worsens macro-F1 for both Qwen models (-13.5 for 7B and -3.3 for 32B on basic chart), suggesting that forcing verbalization of chart content that the model has not attended to yields unreliable descriptions and hallucination. Qualitative analysis of model responses further corroborates the failure patterns (see Appendix F). InternVL3-8B benefits from CoT ($+4.6$ macro-F1), suggesting that model can access and describe chart content but does not draw on it spontaneously when predicting the claim label. The two failure modes are distinct in attention mechanism and response behavior, as observed in the opposite CoT effects across model families (Table 4), indicating that the same intervention is unlikely to be equally effective for both.

5 Conclusion

We investigated why multimodal LLMs perform substantially worse on chart evidence than on table evidence in scientific claim verification, despite the identical underlying data. Through layer-wise probing and attention analysis on three open-weight models, we find that chart-relevant signal is encoded and recoverable from intermediate representations but does not reach the prediction position, indicating a routing failure rather than a perceptual one. Attention analysis further reveals two distinct failure modes. Qwen-family models exhibit an *attention routing failure*, largely bypassing chart evidence when predicting the claim label, while InternVL3-8B shows a *post-attention integration failure*, attending to chart evidence but not integrating it into the prediction. Notably, the two model families with fundamentally different internal failures both show substantially lower accuracy on chart than on table evidence. Output evaluation alone cannot surface this distinction, underscoring that diagnosing format sensitivity requires examining internal computation, not only task performance.

6 Limitations

Our study has the following limitations: First, we use SciTabAlign+ dataset, which contains only 162 unique claims. While this is relatively small, the controlled design, where identical underlying data appears as tables and charts, enables us to isolate format effects that larger uncontrolled datasets cannot capture. The consistency of our findings across model families and chart variants provides confidence in the main result, despite the dataset size. The reported values are taken at the best layer per condition, selected on the test set, which can optimistically bias absolute values. The contrast between mean-pool and last-token decodability on chart evidence, however, holds at every layer.

Second, we analyze only open-weight multimodal LLMs because our analysis requires access to layer-wise hidden states and attention weights. This excludes proprietary models that are widely used in practice. We cannot determine whether the same internal routing failures appear in closed-source models.

Third, our analyses are diagnostic rather than causal. Linear probing and attention patterns provide correlational evidence of how information flows through models, but do not prove models causally use chart information during prediction. We demonstrate that chart information is recoverable from representations but is not reliably reflected in the model's final verification decision.

Finally, we focus on binary scientific claim verification task. It remains unclear whether similar format-specific failures appear in other tasks. Future work should test generalization beyond this setting.

Despite these limitations, our controlled setting makes the main finding informative. Even when underlying evidence data remains unchanged, chart-relevant information can be recoverable from model representations while still failing to support the model's final verification decision.

7 Ethical Considerations

This study uses only publicly available models and dataset, distributed under their respective licenses for research use, and involves no human subjects or new data collection. Our findings highlight that format-sensitive failures in multimodal LLMs may not be visible from model outputs alone. Researchers and practitioners deploying such models should treat model predictions with caution and use diagnostic analyses when evaluating systems for high-stakes scientific workflows. We used AI tools to assist with editing and coding in the preparation of this paper. All content has been thoroughly reviewed and verified by the human authors.

8 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24K03231, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 554559555, and the German Academic Exchange Service (DAAD) - 57557629.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Lianghan Dong and Anamaria Crisan. 2026. Probing the visualization literacy of vision language models: The good, the bad, and the ugly. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):1175–1185.

Parsa Esmaeilkhani and Longin Jan Latecki. 2026. Direct visual grounding by directing attention of visual tokens. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5787–5797.

Xanh Ho, Sunisth Kumar, Yun-Ang Wu, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. 2025. Table-text alignment: Explaining claim verification against tables in scientific papers. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2509–2517, Suzhou, China. Association for Computational Linguistics.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. 2026a. Format matters: The robustness of multimodal llms in reviewing evidence from tables and charts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31014–31022.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, Andre Greiner-Petter, and Akiko Aizawa. 2026b. Sciclaimeval: Cross-modal claim verification in scientific papers. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11060–11071, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Haruka Kawasaki, Ryota Tanaka, and Kyosuke Nishida. 2026. Responses fall short of understanding: Revealing the gap between internal representations and responses in visual document understanding. *Preprint*, arXiv:2604.04411.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yash Kumar Lal, Manikanta Bandham, Mohammad Saqib Hasan, Apoorva Kashi, Mahnaz Koupaee, and Niranjan Balasubramanian. 2025. MuSciClaims: Multimodal scientific claim verification. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3285–3307, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. 2025a. On the perception bottleneck of VLMs for chart understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10829–10841, Suzhou, China. Association for Computational Linguistics.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. 2025b. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *Preprint*, arXiv:2510.17771.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Ahmed Masry, Do Xuan Long, 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.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

et al. Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. 2025. SciVer: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8562–8579, Vienna, Austria. Association for Computational Linguistics.

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. 2024. 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.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19781–19791.

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 Experimental Details

All experiments were conducted on NVIDIA A100 (80GB) GPUs. Qwen2.5-VL-32B was distributed across 2 GPUs, while the 7B and 8B models each ran on a single GPU. Linear probes were implemented in PyTorch and trained from zero initialization on standardized features. Hidden states and attention weights were extracted using the Hugging Face Transformers library (Thomas Wolf, 2020). Under 1000 label permutations, the probe accuracy stays at chance (0.500 ± 0.047, maximum 0.660), far below the observed accuracy ($p < 0.001$). An ℓ_2-regularized scikit-learn (Pedregosa et al., 2011) probe changes accuracy by at most 4.3 points and leaves all comparisons unchanged.

B Layer-wise Probing Analysis - All Models

B.1 Baseline Model Performance

Table 2 shows baseline model performance. Models perform substantially better on tables than on charts, despite identical underlying data.

	Qwen2.5-VL-7B		Qwen2.5-VL-32B		InternVL3-8B	
Format	Acc	F1	Acc	F1	Acc	F1
Table	**67.5**	**70.5**	**80.9**	**83.8**	**69.1**	**70.4**
Basic chart	66.0	65.9	68.5	68.2	56.2	55.2
Chart symbol	59.9	58.7	61.1	59.6	54.9	50.1
Line chart	56.2	55.5	61.1	60.0	56.2	53.0
Swapped chart	56.8	56.7	69.1	68.8	58.0	56.9

Table 2: Model inference accuracy and macro-F1 (%) on SciTabAlign+. Results use the baseline prompt tempate shown in Figure 5 .

Baseline Prompt Template

System: You are an expert in claim verification against scientific papers.

User: [image]
Use the provided image, predict the label for this claim: '{claim}'; the label can be Supported or Refuted. Think step by step before answering. Please format your final answer within brackets as follows: `<ans> YOUR ANSWER </ans>`

Figure 5: Baseline prompt template for claim verification. For the table format, [image] placeholder is replaced with [table] and the table text is prepended as `Table Information: {table_text}`. .

B.2 AUROC Divergence

As shown in Figure 6, table evidence rises sharply in final layers under last-token probing while chart variants decline towards chance. Under mean-pool probing, chart-relevant signal remains decodable across all layers while tables perform worse than charts. Figure 7 visualizes this pattern for last-token probe as a heatmap.

B.3 Image-Token Attention Distribution

InternVL3-8B maintains near-proportional aggregate image-token attention (93%), unlike Qwen models (4–11%). We analyze whether this reflects distributed routing or concentration in summary tokens. Image-token attention in InternVL3-8B is distributed across positions rather than concentrated in summary tokens (Gini = 0.18, CV = 0.41 across 10 random claims). Figure 8 shows that within-image attention concentration drops from 0.71 in early layers to < 0.15 in deeper layers, confirming distributed visual routing. Notably, this near-proportional image attention does not translate to better chart accuracy. InternVL3-8B achieved lower chart accuracy than both Qwen models, as shown in Table 2, indicating that attending to chart evidence does not guarantee effective use of its content. Table 3 compares this distribution across models. The Qwen models show the opposite pattern. The bottom 90% of image tokens receive only 25–34% of image attention, concentrated near the start of the image token sequence.

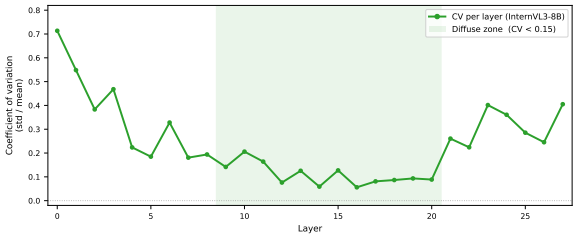


Figure 8: Coefficient of variation in image-token attention across InternVL3-8B layers. Attention diffuses from concentrated (early) to distributed (deeper layers), ruling out summary-token bottlenecks.

Model	Bottom-90% share	Gini
Qwen2.5-VL-7B	25–34%	0.84
Qwen2.5-VL-32B	25–34%	0.79
InternVL3-8B	80–82%	0.18

Table 3: Distribution of attention within image tokens, over the same claims as Figure 8.

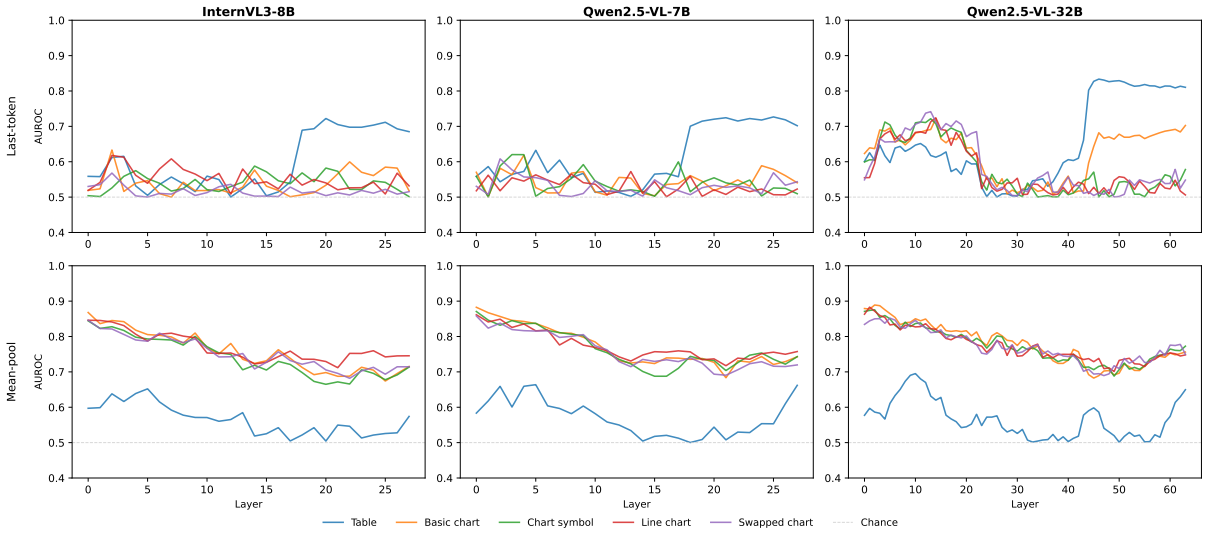


Figure 6: Probe AUROC by layer for all three models on SciTabAlign+. Top row: last-token probing. Bottom row: mean-pool probing. In all models, table evidence (blue) rises sharply in later layers under last-token probing while chart conditions remain near chance over the same range. Under mean-pool probing, chart conditions are consistently decodable across all layers, confirming that chart-relevant signal is present in the model but does not reach the prediction position.

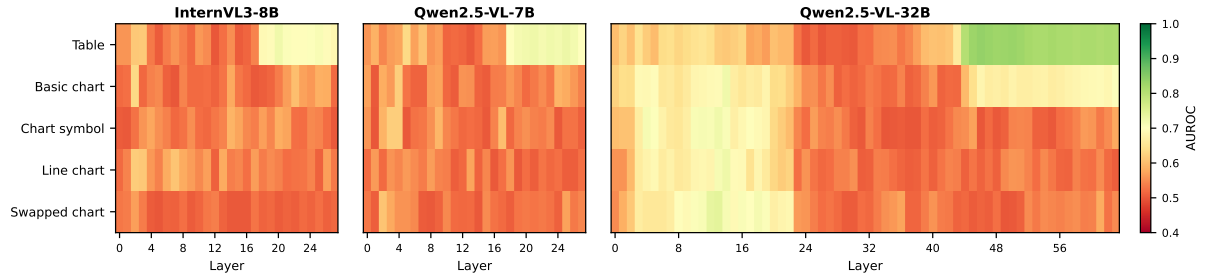


Figure 7: Last-token probe AUROC heatmap across layers and formats for all models. Table evidence (top row) accumulates to high AUROC in deeper layers and chart variants remain near chance across all three models.

C Cross-format Representation Convergence

As shown in Figure 11, table and chart representations converge to high cosine similarity in deeper layers across all models.

	Standard		CoT		Δ	
Model	Acc	F1	Acc	F1	ΔAcc	ΔF1
Qwen2.5-VL-7B	66.0	65.9	45.7	52.4	−20.3	−13.5
Qwen2.5-VL-32B	68.5	68.2	63.6	64.9	−4.9	−3.3
InternVL3-8B	56.2	55.2	59.9	59.8	+3.7	+4.6

Table 4: Effect of chain-of-thought prompting on basic chart accuracy and macro-F1 (%) on SciTabAlign+. Δ values are CoT minus standard.

D Chain-of-Thought Ablation

The prompt template is shown in Figure 9. As shown in Table 4, chain-of-thought prompting degrades performance for Qwen models while improving InternVL3-8B.

E Claim-Only Baseline

Table 5 shows claim-only probes reach 58–67% AUROC vs. 84–89% for charts. This gap shows that the mean-pool probe captures visual information, and not text leakage.

Chain-of-Thought Prompt Template

System: You are an expert in claim verification against scientific papers.

User: [image]
Step 1: Carefully examine the chart. Describe the data you observe — list specific values, comparisons between groups, and any trends visible. Be precise and mention exact numbers shown on axes or labels if you can read them.

Step 2: Based ONLY on what you described in Step 1, determine whether the claim is Supported or Refuted.

Claim: `{claim}`

Think step by step. Format your final answer as: `<ans>` `YOUR ANSWER` `</ans>`

Figure 9: Chain-of-thought prompt template for claim verification against chart format.

Model	Claim-Only	Basic Chart	Δ
Qwen2.5-VL-7B	58.5%	88.3%	+29.8%
Qwen2.5-VL-32B	67.4%	88.9%	+21.5%
InternVL3-8B	62.0%	86.8%	+24.8%

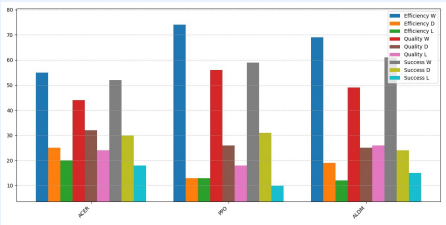
Table 5: Claim-only vs. basic chart mean-pool AUROC. Claim-only probe is trained only on claim text (no image tokens provided). The large gap (Δ) confirms that probe performance is driven by visual information, not just claim-text cues.

F Error Analysis

We randomly selected 5 instances each from cases where the model fails but probe succeeds on basic chart evidence, and where models succeeds but probe fails on table evidence for Qwen2.5-VL-7B, and examined the model responses qualitatively. Chart failures are predominantly driven by hallucinated content, where the model fabricates values or entities not present in the chart. Cases where the model outperforms the probe on table evidence are dominated by negative claims, where the model correctly resolves negation while the probe predicts incorrectly. Figure 10 shows a representative example.

Claim: "Among all the baselines, GDPL does not obtain the most preference against PPO." **True Label:** *Refuted*

Chart Evidence × Model ✓ Probe



Caption: The count of human preference on dialog session pairs that GDPL wins (W), draws with (D) or loses to (L) other methods based on different criteria. One method wins the other if the majority prefer the former one.

Model: × Supported

Probe: ✓ Refuted

Model response (excerpt):

> "GDPL and PPO have the same highest value in the 'Success W' category, neither is definitively preferred over the other [. . .] the claim is supported."

The model hallucinates GDPL's presence in the chart despite it not appearing as an explicit bar. The mean-pool probe correctly decodes the *Refuted* label.

Table Evidence ✓ Model × Probe

VS.	Efficiency			Quality			Success		
	W	D	L	W	D	L	W	D	L
ACER	55	25	20	44	32	24	52	30	18
PPO	74	13	13	56	26	18	59	31	10
ALDM	69	19	12	49	25	26	61	24	15

Caption: The count of human preference on dialog session pairs that GDPL wins (W), draws with (D) or loses to (L) other methods based on different criteria. One method wins the other if the majority prefer the former one.

Model: ✓ Refuted

Probe: × Supported

Model response (excerpt):

> "GDPL wins 55 times and loses 13 times in efficiency [. . .] GDPL wins more often than it loses in all three criteria, the claim is refuted."

The model correctly reads the structured values and resolves double negation in the claim, and makes correct prediction. The probe predicts *Supported* incorrectly, consistent with the pattern observed across negative claims in the samples instances for this analysis.

Figure 10: Representative error analysis example for Qwen2.5-VL-7B. **Left:** Chart evidence where the model fails through hallucination while the probe succeeds. **Right:** Table evidence where the model succeeds through negation handling while the probe fails.

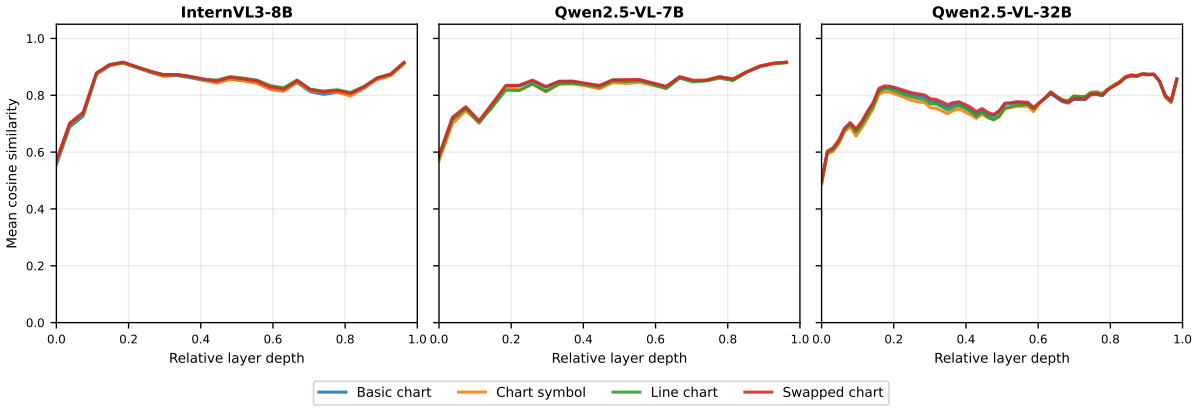


Figure 11: Mean cosine similarity between table and chart mean-pooled representations across all layers for each model (excluding final LM head layer). Table and chart representations start with low similarity in early layers but converge closer in deeper layers across all models and chart variants.