SciLens: Multi-modal Scientific Claim Verification with Agentic Entailment and Grounding

Anonymous Author(s)

Abstract

Scientific discovery increasingly relies on automated systems that generate hypotheses, inspect multimodal evidence, and validate claims at scale. Yet scientific claim verification is not well served by asking a vision-language model for a direct binary judgment: claims often combine numerical results, comparisons, scope qualifiers, and explanatory context, while evidence is encoded in tables and figures with distinct grounding structures. We present **SciLens**, an evidence-conditioned atomic entailment framework for multimodal scientific claim verification. SciLens decomposes each claim into central empirical atoms and background atoms, grounds the central atoms to modality-specific evidence witnesses, and predicts the final label with an atom-level entailment rule. For tables, atoms are grounded to rows, columns, cells, arithmetic relations, and table scope; for figures, they are grounded through panels, axes, legends, visual encodings, categories, trends, ranks, and qualifier checks. This yields a unified validation procedure in which a claim is supported only if every central empirical atom is entailed by the current evidence. On the SciClaimEval development set, SciLens achieves 79.2% macro-F1 and 63.1% pair accuracy, showing that structured agentic validation improves both evidence sensitivity and interpretability.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Information extraction**; *Computer vision tasks.*

Keywords

Scientific claim verification, multimodal reasoning, agentic workflow, evidence grounding, entailment

ACM Reference Format:

Anonymous Author(s). 2026. SciLens: Multi-modal Scientific Claim Verification with Agentic Entailment and Grounding. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Agentic AI systems are increasingly used in scientific workflows, where language models are combined with planning, tool use, retrieval, and multimodal reasoning [7, 18]. However, scientific progress depends not only on generating hypotheses or fluent summaries, but also on verifying whether reported claims are supported by the evidence in a paper. This problem is challenging because evidence is often encoded in dense tables and multi-panel figures, while claims combine exact values, comparisons, scope qualifiers, and explanatory context.

SciClaimEval [5] provides a focused benchmark for this setting. A system is given one scientific claim and one evidence item, either a table or a figure, and must predict whether the evidence supports or refutes the claim. Unlike generic visual question answering, this task requires evidence-sensitive verification: a model must locate the relevant row, column, panel, axis, legend entry, category, trend, rank, or statistical cue, rather than relying on the caption or the general plausibility of the claim.

We present SciLens, an agentic framework that formulates scientific table and figure claim verification as a unified evidence-conditioned entailment problem. SciLens decomposes each claim into checkable empirical atoms, grounds those atoms to modality-specific evidence witnesses, and predicts Supported only when the current table or figure entails the central empirical content of the claim. The framework is lightweight and task-aligned: it does not retrieve external documents, compare against paired examples, or use mutation metadata. Instead, it reasons only over the given claim, the current evidence item, and optional local context provided with the sample.

On the SciClaimEval development set, SciLens achieves 79.2% macro-F1 and 63.1% pair accuracy, improving over the strongest vanilla Qwen3-VL-30B-A3B baseline by 3.2 and 8.3 points, respectively. The gains are especially large for pair accuracy, suggesting that atom-level grounding improves evidence sensitivity rather than merely increasing claim-level plausibility. These results show that structured agentic validation can provide an effective and interpretable path toward multimodal scientific claim verification.

2 The SciLens Framework

2.1 Problem Formulation

Given a claim c, an evidence item e, optional local context x, and modality $m \in \{\text{table}, \text{figure}\}$, SciClaimEval Subtask 1 requires a binary prediction $y \in \{\text{Supported}, \text{Refuted}\}$. We formulate the task as evidence-conditioned entailment:

$$\hat{y} = f_\theta(c, e, x, m).$$

The prediction must be entailed by the current evidence item, rather than by claim plausibility, caption-only cues, or evidence elsewhere in the paper. SciLens operationalizes this principle through three steps: decompose the claim into empirical atoms, ground each atom to modality-specific evidence witnesses, and predict Supported only when all central atoms are verified.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

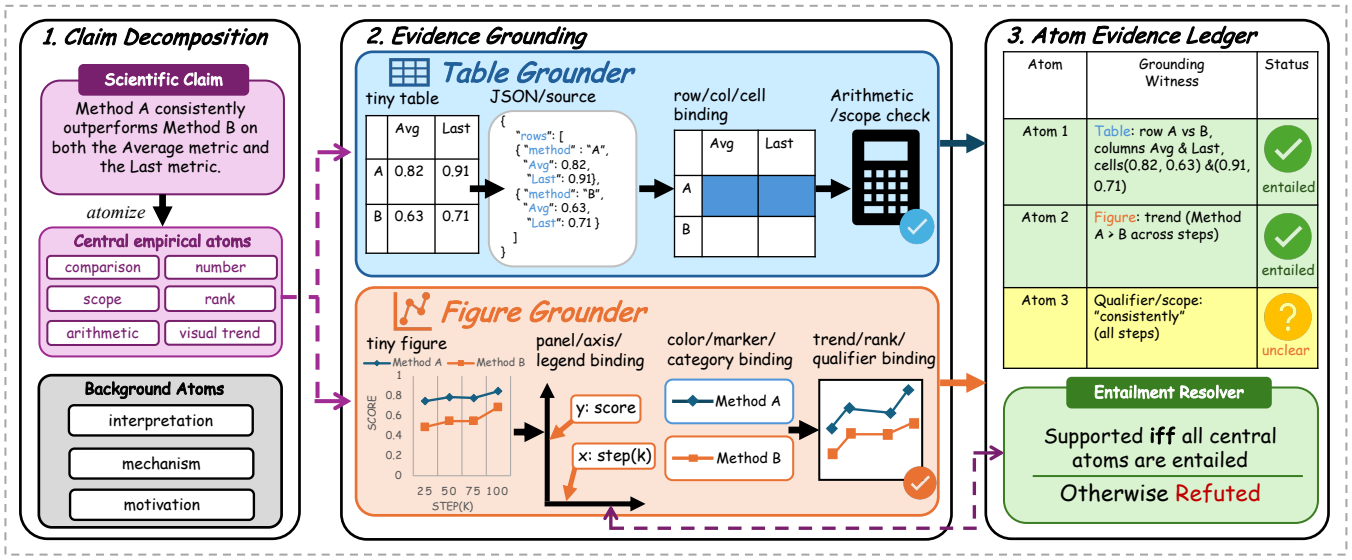

Figure 1: Overview of the SciLens framework.

2.2 Evidence Pack Construction

For each instance, SciLens constructs an evidence pack containing the claim, modality, caption, optional paragraph context, and evidence image. For tables, the pack also includes structured table JSON and source-table text when available, which help recover headers, cells, units, signs, and OCR-sensitive content. These auxiliary representations are used only to interpret the current table. For figures, the image remains the primary evidence, while caption and context help identify panels, metrics, or domain conventions without overriding the displayed visual content.

The pack is intentionally isolated: it excludes paired evidence items, gold labels, mutation operations, and previous predictions. This prevents pair-aware shortcuts and matches the evaluation protocol, where each instance is classified independently.

2.3 Claim Decomposition

The first stage decomposes the claim into central empirical atoms and background atoms:

$$D(c) = A_{\text{central}}(c) \cup A_{\text{bg}}(c).$$

Central atoms are label-determining requirements, such as entities, metrics, values, comparison directions, ranks, counts, arithmetic relations, and statistical qualifiers. Background atoms capture mechanisms, motivations, or explanatory statements that need not be proven unless directly contradicted by the evidence. This separation reflects scientific writing conventions: a table may establish that a method ranks first without verifying the proposed explanation for why it does so. The resolver therefore focuses on the empirical core while avoiding over-penalization of explanatory phrasing.

2.4 Table Grounding

For table evidence, SciLens uses a three-stage workflow: claim decomposition, evidence grounding, and entailment resolution. The table grounder aligns each central atom to concrete witnesses, including row labels, column headers, cells, units, signs, formatting cues, table scope, and arithmetic. For claims involving ratios, averages, differences, percentage points, fold changes, counts, or significance statements, the grounder records the calculation or table cue used for verification. We represent the witness for atom a_i as

$$w_i^{\text{tab}} = (r_i, h_i, v_i, u_i),$$

where r_i, h_i, v_i, and u_i denote the row, header, cell value, and unit.

The table verifier treats exact numbers, named metrics, row–column bindings, ranks, and hard scope terms such as "all", "only", and "every" as strict constraints. Approximate phrases such as "around", "similar", or "slightly" are accepted only when both direction and magnitude are compatible with the table. This policy preserves ordinary scientific approximations while enforcing truth-changing constraints.

2.5 Figure Grounding

Figure evidence requires visual rather than tabular witnesses. SciLens therefore uses a reader–auditor–resolver workflow. The reader extracts claim-critical visual requirements, including the relevant panel, metric, axis direction, category relation, comparison, trend, rank, peak, numeric relation, and qualifier. The auditor then binds entities and groups to visual encodings in the current image, such as legends, colors, markers, line styles, labels, panels, axes, and color bars, before assessing values or trends. For figure atom a_i, we denote the binding as

$$B_i^{\text{fig}} = \{(q, \ell_q) : q \in Q_i\},$$

where Q_i is the set of claim-critical visual slots and ℓ_q is the image element assigned to slot q.

This binding step targets common failure modes in scientific figures: swapped legends, incorrect panels, flipped axes, altered categories, wrong trend directions, and missing significance markers.

Visual mappings mentioned in the claim or caption are treated as hypotheses to verify against the image, not as facts. If a central visual binding is missing or unreadable, SciLens predicts Refuted unless another explicit visual cue verifies the relation.

2.6 Typed Qualifier Policy

SciLens applies a shared qualifier policy across modalities. Hard logical qualifiers, such as "all", "every", "only", "highest", "lowest", "best", and "same", are checked strictly because a single counterexample can falsify the claim. Soft descriptive qualifiers, such as "usually", "slightly", "similar", and "more pronounced", are judged by the dominant tabular or visual trend rather than by a fixed threshold. Statistical qualifiers, including significance, p-values, confidence intervals, and error bars, are central when explicitly invoked by the claim and otherwise serve only as compatible supporting cues.

2.7 Entailment Resolution and Guards

The final resolver takes the evidence pack and same-instance stage notes and outputs a JSON verdict with a short rationale. The decision rule is conservative:

$$\text{Supported}(c, e) \iff \forall a_i \in A_{\text{central}}(c),\ e \models a_i.$$

Thus, SciLens predicts Refuted if any central atom is contradicted, missing, unclear, misaligned, or supported only by caption/context. This atom-level rule yields an interpretable trace by identifying the decisive evidence requirement rather than returning only a binary label.

We add lightweight consistency guards to prevent later stages from overriding grounded evidence. For tables, a Refuted prediction is preserved when a central atom has an explicit contradiction; otherwise unsupported refutations may be revised to Supported. For figures, the resolver is invoked when earlier stages disagree or report uncertainty, and explicit auditor contradictions are preserved against generic supported flips. These guards are heuristic but use only same-instance traces, avoiding leakage from labels, paired evidence, or mutation metadata.

3 Experiments

3.1 Dataset

We evaluate on the SciClaimEval Subtask 1 development set, which contains 747 claim–evidence instances: 482 table instances and 265 figure instances. Each instance consists of a scientific claim, a table or figure evidence image, a caption, optional local context, and a binary label $y \in \{\text{Supported}, \text{Refuted}\}$. The split contains 395 Supported and 352 Refuted instances.

SciClaimEval also groups many instances into paired examples through `claim_id_pair`. Pair accuracy counts a group as correct only if all predictions in that group are correct. Following the official evaluation setting, instances marked as `no pair` are excluded from this metric, yielding 236 table pairs and 116 figure pairs.

3.2 Models and Settings

We compare SciLens against vanilla multimodal LLM baselines. In the vanilla setting, a single prompt asks the model to inspect the claim and the current evidence item and output a JSON label. In the agentic setting, the same prediction is produced by the staged framework in section 2, using only same-instance evidence and intermediate stage notes.

Our main backbones are Qwen3-VL-8B and Qwen3-VL-30B-A3B [1]. For table instances, SciLens uses three calls: claim decomposition, table grounding, and entailment resolution. Structured table JSON and source-table text are provided when available as reading aids. For figure instances, SciLens uses a reader–auditor–resolver workflow over the current figure image and derived image views. Unless otherwise noted, we use temperature 0.25 for tables, temperature 0.6 for figures, top-$p = 0.95$, top-$k = 20$, and seed 1234. Outputs are normalized to Supported or Refuted. We report macro-F1 over individual instances and pair accuracy over paired groups.

3.3 Results and Analysis

Aggregate results. Table 1 reports results for table evidence, figure evidence, and the full development set. Across both backbones, SciLens improves over direct multimodal prompting. With Qwen3-VL-30B-A3B, SciLens achieves 79.2 macro-F1 and 63.1 pair accuracy overall, improving over the vanilla baseline by 3.2 and 8.3 points, respectively. With Qwen3-VL-8B, SciLens improves macro-F1 by 4.1 points and pair accuracy by 8.5 points. The larger gain in pair accuracy suggests that staged grounding improves evidence sensitivity, since paired examples require the model to distinguish supporting from refuting evidence for the same claim.

Performance differs by modality. Table verification is stronger overall: with Qwen3-VL-30B-A3B, SciLens reaches 81.7 macro-F1 and 67.4 pair accuracy on table evidence, reflecting the benefit of structured table representations and row–column–cell grounding. Figure verification remains harder but benefits substantially from the agentic workflow: figure pair accuracy with the 30B backbone rises from 38.8 to 54.3, indicating that explicit checks over panels, legends, categories, and trends help reduce brittle direct judgments.

Efficiency. The two backbones show different generation costs under the same agentic procedure. In our raw stage logs, Qwen3-VL-8B generates about 52.9K tokens per table instance and 38.1K per figure instance, whereas Qwen3-VL-30B-A3B generates about 26.5K and 23.2K tokens, respectively. Since the prompt-side evidence packs are comparable, the difference mainly comes from intermediate reasoning. This suggests that the larger backbone executes the staged decomposition more compactly and uses the evidence more efficiently, especially for figure verification.

Operation-level results. Figure 2 breaks down pair accuracy by evidence modification type. For tables, SciLens improves most on changed cell values: pair accuracy increases from 53.0 to 64.5 with Qwen3-VL-8B and from 62.8 to 69.9 with Qwen3-VL-30B-A3B. Gains are smaller for swapped rows or columns: the 8B agent improves from 47.2 to 52.8, while the 30B agent drops from 62.3 to 58.5. This suggests that altered numerical entries are well handled by row–column–cell grounding, but row/column alignment remains difficult.

For figures, the largest gains occur when verification requires explicit visual binding. With Qwen3-VL-30B-A3B, graph-flip pair accuracy rises from 35.0 to 70.0, legend-swap accuracy rises from 39.7 to 54.0, and graph-swap examples improve by 10.0 points for both backbones. However, category swaps remain challenging:

Table 1: Main results on the SciClaimEval development set. Macro-F1 is computed over individual claim–evidence samples, while pair accuracy is computed over paired evidence groups. All values are percentages; n denotes the number of individual samples and p denotes the number of paired evidence groups.

Models	Table		Figure		Overall	
	F1 $n = 482$	Pair Acc. $p = 236$	F1 $n = 265$	Pair Acc. $p = 116$	F1 $n = 747$	Pair Acc. $p = 352$
Vanilla LLM (baseline)						
InternVL3_5-8B	70.6	44.5	52.2	16.4	65.4	35.2
InternVL3_5-14B	72.0	49.2	59.7	25.9	68.5	41.5
InternVL3_5-38B	71.7	47.5	57.9	25.0	67.8	40.1
Qwen3-VL-4B	74.2	49.2	63.7	32.8	70.6	43.8
Qwen3-VL-8B	75.4	51.7	65.8	37.1	72.1	46.9
Qwen3-VL-30B-A3B	80.6	62.7	66.6	38.8	76.0	54.8
SciLens Agent						
Qwen3-VL-8B	80.5 (+5.1)	61.9 (+10.2)	67.7 (+1.9)	42.2 (+5.1)	76.2 (+4.1)	55.4 (+8.5)
Qwen3-VL-30B-A3B	**81.7** (+1.1)	**67.4** (+4.7)	**74.3** (+7.7)	**54.3** (+15.5)	**79.2** (+3.2)	**63.1** (+8.3)

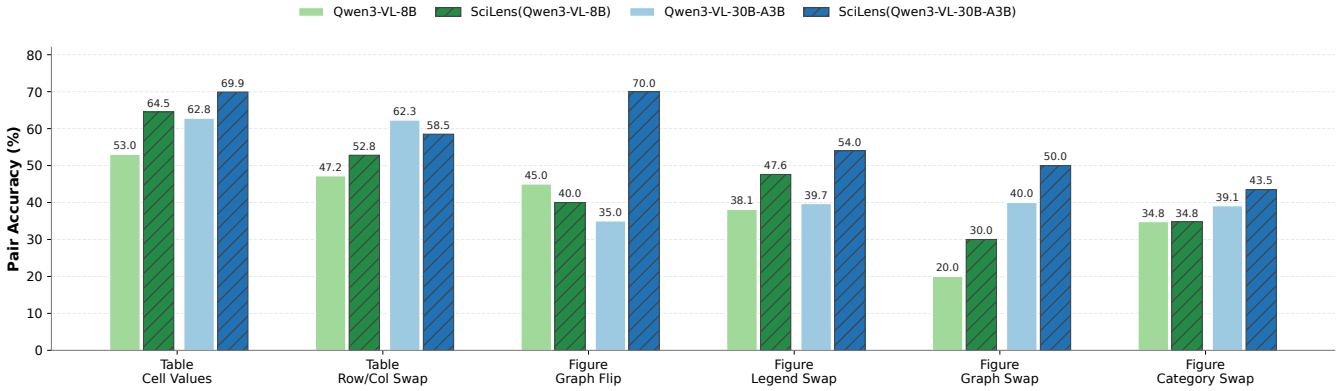

Figure 2: Operation-level comparison between SciLens agents and vanilla Qwen3-VL baselines.

accuracy is unchanged for 8B at 34.8 and improves only from 39.1 to 43.5 for 30B. These results show that SciLens improves binding to axes, legends, panels, and graph identity, while fine-grained category matching and directional interpretation remain brittle.

4 Related Works

4.1 Entailment Verification

Entailment verification in the LLM era has shifted from static sentence-pair classification to retrieval-augmented, rationale-aware, and evidence-grounded reasoning over complex claims [12, 13]. Recent work studies how LLMs perform structured logical inference, proof construction, and entailment verification under diverse scenarios [4, 16, 17, 20]. Parallel efforts extend verification beyond text by requiring models to reason over visual evidence, tables, figures, and multimodal contexts [2, 6, 8].

4.2 LLM Agents in Scientific Discovery

LLM-based scientific agents are increasingly involved in scientific discovery throughout the research lifecycle [7, 18]. Recent studies further push scientific discovery agents toward deeper autonomy, for tasks ranging from scientific deep research to scientific law discovery [3, 10, 15, 19, 21]. With the development of multi-modality in agentic systems, long-context vision-language modeling and persistent multimodal memory have become important for analyzing scientific documents and maintaining grounded evidence across extended trajectories [9, 11, 14].

5 Conclusion

We presented SciLens, an agentic framework for scientific claim verification with atomic claim decomposition and modality-specific evidence grounding. On the SciClaimEval development set, SciLens improves over vanilla multimodal prompting, achieving 79.2 macro-F1 and 63.1 pair accuracy with Qwen3-VL-30B-A3B. The strongest gains occur in pair accuracy and visually grounded figure cases, suggesting improved evidence sensitivity, while remaining errors on row/column swaps, fine-grained category matching, and directional visual reasoning highlight directions for future work.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] https://arxiv.org/abs/2511.21631

[2] Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-Tuple-Table: Towards Information Integration in Text-to-Table Generation via Global Tuple Extraction. arXiv:2404.14215 [cs.CL] https://arxiv.org/abs/2404.14215

[3] Tianqing Fang, Zhisong Zhang, Xiaoyang Wang, Rui Wang, Can Qin, Yuxuan Wan, Jun-Yu Ma, Ce Zhang, Jiaqi Chen, Xiyun Li, Yonglin Wang, Jingchen Ni, Tianshi Zheng, Chun Chen, Wenhao Yu, Zhenwen Liang, Hongming Zhang, Haitao Mi, and Dong Yu. 2026. Cognitive Kernel-Pro: A Framework for Deep Research Agents and Agent Foundation Models Training. arXiv:2508.00414 [cs.AI] https://arxiv.org/abs/2508.00414

[4] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alex Wardle-Solano, Hannah Szabo, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander R. Fabbri, Wojciech Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: Natural Language Reasoning with First-Order Logic. arXiv:2209.00840 [cs.CL] https://arxiv.org/abs/2209.00840

[5] Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, Andre Greiner-Petter, and Akiko Aizawa. 2026. SciClaimEval: Cross-modal Claim Verification in Scientific Papers. arXiv:2602.07621 [cs.CL] https://arxiv.org/abs/2602.07621

[6] Fangzhou Liang, Tianshi Zheng, Chunkit Chan, Yauwai Yim, and Yangqiu Song. 2025. LLM-Hanabi: Evaluating Multi-Agent Gameplays with Theory-of-Mind and Rationale Inference in Imperfect Information Collaboration Game. arXiv:2510.04980 [cs.AI] https://arxiv.org/abs/2510.04980

[7] Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research. arXiv:2501.04306 [cs.CL] https://arxiv.org/abs/2501.04306

[8] Yunxiang Mo, Tianshi Zheng, Qing Zong, Jiayu Liu, Baixuan Xu, Yauwai Yim, Chunkit Chan, Jiaxin Bai, and Yangqiu Song. 2025. DixitWorld: Evaluating Multimodal Abductive Reasoning in Vision-Language Models with Multi-Agent Dixit Gameplay. arXiv:2510.10117 [cs.AI] https://arxiv.org/abs/2510.10117

[9] Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, Zhengnan Li, Yangqiu Song, Ginny Wong, and Simon See. 2026. MemLens: Benchmarking Multimodal Long-Term Memory in Large Vision-Language Models. arXiv:2605.14906 [cs.CV] https://arxiv.org/abs/2605.14906

[10] Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. 2025. LLM-SR: Scientific Equation Discovery via Programming with Large Language Models. arXiv:2404.18400 [cs.LG] https://arxiv.org/abs/2404.18400

[11] Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, Jianing Wang, Qintong Li, Xiangru Tang, Tianbao Xie, Xiachong Feng, Xiang Li, Ben Kao, Wenhai Wang, Biqing Qi, Lingpeng Kong, and Zhiyong Wu. 2026. ScienceBoard: Evaluating Multimodal Autonomous Agents in Realistic Scientific Workflows. arXiv:2505.19897 [cs.AI] https://arxiv.org/abs/2505.19897

[12] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. arXiv:1803.05355 [cs.CL] https://arxiv.org/abs/1803.05355

[13] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. arXiv:2210.13777 [cs.CL] https://arxiv.org/abs/2210.13777

[14] Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. 2025. MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly. arXiv:2505.10610 [cs.CV] https://arxiv.org/abs/2505.10610

[15] Wanghan Xu, Yuhao Zhou, Yifan Zhou, Qinglong Cao, Shuo Li, Jia Bu, Bo Liu, Yixin Chen, Xuming He, Xiangyu Zhao, Xiang Zhuang, Fengxiang Wang, Zhiwang Zhou, Qiantai Feng, Wenxuan Huang, Jiaqi Wei, Hao Wu, Yuejin Yang, Guangshuai Wang, Sheng Xu, Ziyan Huang, Xinyao Liu, Jiyao Liu, Cheng Tang, Wei Li, Ying Chen, Junzhi Ning, Pengfei Jiang, Chenglong Ma, Ye Du, Changkai Ji, Huihui Xu, Ming Hu, Jiangbin Zheng, Xin Chen, Yucheng Wu, Feifei Jiang, Xi Chen, Xiangru Tang, Yuchen Fu, Yingzhou Lu, Yuanyuan Zhang, Lihao Sun, Chengbo Li, Jinzhe Ma, Wanhao Liu, Yating Liu, Kuo-Cheng Wu, Shengdu Chai, Yizhou Wang, Ouwen Zhangjin, Chen Tang, Shufei Zhang, Wenbo Cao, Junjie Ren, Taoyong Cui, Zhouheng Yao, Juntao Deng, Yijie Sun, Feng Liu, Wangxu Wei, Jingyi Xu, Zhangrui Li, Junchao Gong, Zijie Guo, Zhiyu Yao, Zaoyu Chen, Tianhao Peng, Fangchen Yu, Bo Zhang, Dongzhan Zhou, Shixiang Tang, Jiaheng Liu, Fenghua Ling, Yan Lu, Yuchen Ren, Ben Fei, Zhen Zhao, Xinyu Gu, Rui Su, Xiao-Ming Wu, Weikang Si, Yang Liu, Hao Chen, Xiangchao Yan, Xue Yang, Junchi Yan, Jiamin Wu, Qihao Zheng, Chenhui Li, Zhiqiang Gao, Hao Kong, Junjun He, Mao Su, Tianfan Fu, Peng Ye, Chunfeng Song, Nanqing Dong, Yuqiang Li, Huazhu Fu, Siqi Sun, Lijing Cheng, Jintai Lin, Wanli Ouyang, Bowen Zhou, Wenlong Zhang, and Lei Bai. 2025. Probing Scientific General Intelligence of LLMs with Scientist-Aligned Workflows. arXiv:2512.16969 [cs.AI] https://arxiv.org/abs/2512.16969

[16] Tianshi Zheng, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025. The Curse of CoT: On the Limitations of Chain-of-Thought in In-Context Learning. arXiv:2504.05081 [cs.CL] https://arxiv.org/abs/2504.05081

[17] Tianshi Zheng, Jiayang Cheng, Chunyang Li, Haochen Shi, Zihao Wang, Jiaxin Bai, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025. LogiDynamics: Unraveling the Dynamics of Inductive, Abductive and Deductive Logical Inferences in LLM Reasoning. arXiv:2502.11176 [cs.CL] https://arxiv.org/abs/2502.11176

[18] Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. arXiv:2505.13259 [cs.CL] https://arxiv.org/abs/2505.13259

[19] Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K. Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2026. NewtonBench: Benchmarking Generalizable Scientific Law Discovery in LLM Agents. arXiv:2510.07172 [cs.AI] https://arxiv.org/abs/2510.07172

[20] Tianshi Zheng, Jiazheng Wang, Zihao Wang, Jiaxin Bai, Hang Yin, Zheye Deng, Yangqiu Song, and Jianxin Li. 2025. Enhancing Transformers for Generalizable First-Order Logical Entailment. arXiv:2501.00759 [cs.CL] https://arxiv.org/abs/2501.00759

[21] Tianshi Zheng, Rui Wang, Xiyun Li, Yangqiu Song, and Tianqing Fang. 2026. SciResearcher: Scaling Deep Research Agents for Frontier Scientific Reasoning. arXiv:2605.01489 [cs.AI] https://arxiv.org/abs/2605.01489

A Prompt Details

The full prompts used by SciLens are longer and include evidence-specific content such as captions, selected table rows, source snippets, image descriptions, and previous same-item stage notes. We show compact versions of the reusable stage instructions below. All prompts are single-evidence prompts: the model is not allowed to use pair identifiers, modification operations, gold labels, paired evidence, previous predictions, or external knowledge.

A.1 Shared Prompt Contract

Evidence-Conditioned Verification Contract

Input. One scientific claim, one current evidence item, caption, and allowed auxiliary context.
Evidence priority. The current table or figure image is the decisive evidence. Structured table JSON, source table text, caption, context, and paper-focus text may help identify entities, units, panels, or terminology, but they cannot override what is shown in the current evidence.
Decision rule. Predict `Supported` only when every central empirical atom in the claim is entailed by the current evidence. Predict `Refuted` if any central atom is contradicted, missing, unclear, misaligned, or supported only by auxiliary text.

A.2 Table Agent Prompts

Table Round 1: Claim Decomposer

You are the claim-decomposition stage for an independent table verification agent. Do not decide the final label.
Task. Split the claim into *central empirical atoms* and *background atoms*. Central atoms must be table-checkable requirements about entities, rows, columns, values, metrics, ranks, directions, counts, arithmetic, significance, scope, or formatting cues. Background atoms are mechanisms, motivations, interpretations, or definitions that the table need not prove unless the evidence directly contradicts them.
Strictness. Keep exact numbers, named metrics, row/column bindings, rank terms, and hard scope terms strict. Mark approximate or qualitative wording separately so later stages can judge direction and reasonable magnitude.
Return JSON fields. `central_atoms`, `background_atoms`, and `decomposition_notes`. Each central atom includes `atom`, `atom_type`, `strictness`, and `needed_evidence`.

Table Round 2: Evidence Grounder

You are the evidence-grounding stage for one current table item. Use the Round 1 atoms as a checklist.
Task. For each central atom, align it to concrete table evidence: row name, column name, cell value, unit, formatting cue, full-scope inspection, or arithmetic expression. Structured JSON and source text only help read the current table; caption/context cannot prove cell values.
Checks. For exact values and row–column bindings, cite the specific cell. For `all`, `only`, `best`, `highest`, or `consistently`, inspect the relevant full scope. For averages, ratios, differences, or percent-point claims, write the calculation.
Return JSON fields. `atom_checks`, `contradictions`, `missing_or_unclear`, and `grounding_notes`. Each atom check includes `atom`, `status`, `evidence`, and optional `calculation`.

Table Round 3: Entailment Resolver

You are the final table entailment resolver. Use the evidence pack, claim decomposition, and grounding notes for this same evidence item only.
Resolve. `Supported` if all central empirical atoms are entailed by the current table and no central contradiction remains. `Refuted` if any central atom is contradicted, missing, unclear, misaligned, or supported only by caption/context/paper text. Background atoms do not cause `Refuted` unless the table directly contradicts them.
Return JSON fields. `verdict`, `brief_reason`, `decisive_atoms`, and optional `resolver_checks`.

A.3 Figure Agent Prompts

Figure Round 1: Visual Reader

You are the figure reader for an independent scientific claim verification agent. Inspect one claim and one current figure only.
Task. Convert the claim into atomic visual checks. Read the relevant panel, title, axes, axis direction, legends, colors, line styles, markers, categories, values, trends, peaks, and rankings from the current image. Return a provisional label, but prioritize visual facts, contradictions, and uncertainty.
Return JSON fields. `claim_atoms`, `visual_facts`, `contradictions`, `uncertain`, `provisional_label`, and `brief_reason`.

Figure Round 2: Evidence-Binding Auditor

You are the evidence-binding auditor for one current figure item.
Task. Before judging the claim, bind each compared entity through the current image: panel → axis/metric → legend/color/marker/line/label → observed value, trend, rank, peak, or direction. Do not infer colors, categories, rankings, or values from the claim or caption.
Qualifier policy. Hard logical qualifiers such as `all`, `every`, `only`, `highest`, `lowest`, `same`, and `during all` are strict. Statistical qualifiers require visual markers or reported values when the claim explicitly depends on significance or confidence intervals. Soft qualifiers such as `slightly`, `similar`, `better`, and `more pronounced` are judged by visible trend or relative gap.
Return JSON fields. `label`, `brief_reason`, `atom_checks`, `binding_checks`, `qualifier_checks`, `contradictions`, and `uncertain`.

Figure Round 3: Visual Entailment Resolver

You are the final resolver for one current figure item. Use only this claim, this evidence, and previous same-evidence stage notes.
Resolve. Identify the figure-checkable empirical core: visual comparison, trend, rank, peak, direction, category relation, numeric relation, and truth-changing qualifier. Predict `Supported` only if current-image bindings are correct, the core visual assertion is supported, hard qualifiers hold, required statistical evidence is visible when needed, and no central contradiction remains. Predict `Refuted` for binding failures, panel/metric/category mismatches, wrong direction/rank/peak, missing or unreadable central evidence, or failed hard/required qualifiers.
Return JSON fields. `label` and `brief_reason`.