MUSCICLAIMS: Multimodal Scientific Claim Verification

Yash Kumar Lal **Manikanta Bandham** **Mohammad Saqib Hasan**
Apoorva Kashi **Mahnaz Koupaee** **Niranjan Balasubramanian**
Stony Brook University
ylal@cs.stonybrook.edu

Abstract

Assessing scientific claims requires identifying, extracting, and reasoning with multimodal data expressed in information-rich figures in scientific literature. Despite the large body of work in scientific QA, figure captioning, and other multimodal reasoning tasks over chart-based data, there are no readily usable multimodal benchmarks that directly test claim verification abilities. To remedy this gap, we introduce a new benchmark MUSCICLAIMS accompanied by diagnostics tasks. We automatically extract supported claims from scientific articles, which we manually perturb to produce contradicted claims. The perturbations are designed to test for a specific set of claim verification capabilities. We also introduce a suite of diagnostic tasks that help understand model failures. Our results show most vision-language models are poor (∼0.3-0.5 F1), with even the best model only achieving 0.72 F1. They are also biased towards judging claims as supported, likely misunderstanding nuanced perturbations within the claims. Our diagnostics show models are bad at localizing correct evidence within figures, struggle with aggregating information across modalities, and often fail to understand basic components of the figure.[1]

1 Introduction

Scientific claim verification aims to assess the validity and correctness of a claim with respect to given scientific literature (Kotonya and Toni, 2020a; Saakyan et al., 2021; Mohr et al., 2022; Wadden et al., 2020). Existing work on scientific claim verification mainly focuses on textual data. They pose verification tasks over a single article or text snippet (Kotonya and Toni, 2020a; Saakyan et al., 2021; Mohr et al., 2022), a corpus of full-text articles (Wadden et al., 2020), or larger collections of scientific abstracts (Wadden et al., 2022a).

However, scientific evidence is often presented as heterogeneous information-rich figures that support the important findings, claims and conclusions of experiments. Therefore, scientific claim verification requires both textual and visual understanding capabilities. To assess a claim, one has to go over the figure and its caption, find the panel(s) with information relevant to the claim, combine this visual knowledge with textual information in the figure caption, and finally judging whether the claim is supported or not. While there is a large number of benchmarks on scientific figures, they focus on image captioning (Hsu et al., 2021), question answering (Kahou et al., 2017), or other reasoning tasks (Yue et al., 2024a). There are no readily usable multimodal benchmarks for scientific claim verification. The closest work, ChartCheck (Akhtar et al., 2024), poses a multimodal claim verification task but is restricted to simple data charts crawled from the web, which are substantially different from complex figures found in scientific articles.

To address this gap, we introduce MUSCICLAIMS[2], a multimodal benchmark for claim verification over figures in scientific (physics, chemistry and biology) literature. We set two desiderata for our benchmark: the dataset needs carefully constructed claims that are not supported or have contradictory information in the figures; apart from quantifying model performance, the dataset should also be diagnostic in nature to identify specific model weaknesses. Our dataset creation methodology is designed to meet these desiderata.

We extract claims with inline references to figures from the results section of articles. We manually filter these to only retain claims that are clearly and unambiguously supported by the figures. Then, we create contradictory claims by perturbing these supporting claims. We devise a diverse set of per-

[1]Data is available at https://huggingface.co/datasets/StonyBrookNLP/MuSciClaims/

[2]Code is available at https://github.com/StonyBrookNLP/musciclaims

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

turbations to test specific capabilities for claim verification including qualitative and quantitative reasoning, and observation-inference connections.

Last, we create a suite of diagnostic tasks associated with each claim to better understand model failures. Specifically, we design tasks that help uncover errors across aspects of basic visual understanding, evidence localization, cross-modal aggregation, and epistemic sensitivity. We ensure the integrity of the dataset through manual analysis. The resulting dataset consists of 1515 (`claim`, `figure`) data points from PHYSICS, CHEMISTRY and BIOLOGY, equally balanced across 3 class labels (SUPPORT, NEUTRAL, CONTRADICT), each accompanied by diagnostic questions.

We benchmark a suite of visual language models (VLMs) on MUSCICLAIMS. Most models are poor at scientific claim verification out-of-the-box. Prompting VLMs to explain their decisions helps performance, but only slightly. Despite these gains, there is still a large room for improvement. Our diagnostics shows that models fail at evidence localization, introducing noise in their reasoning process consequently performing worse. Their basic visual understanding and cross-modal aggregation capabilities also need improvement.

In summary, our contributions are:

1. We present MUSCICLAIMS, an evaluation benchmark for multimodal scientific claim verification over information-rich figures.
2. We find that contemporary models are good, but have significant room for improvement on claim verification.
3. Our diagnostic tests pinpoint specific model abilities to improve—localizing to the right information and cross-modal information aggregation—for better claim verification.

2 Related Work

Multimodal Scientific Benchmarks There has been extensive work on evaluating multimodal understanding abilities of contemporary models. Some work focuses on image captioning tasks where, given an image, the model is asked to generate a concise description for it (Hsu et al., 2021; Tang et al., 2023). But the larger share belongs to question answering benchmarks. These benchmarks differ on types of image, questions, knowledge required to answer questions, domains, scale, and annotations. While FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018) and PlotQA (Methani et al., 2020) provide large-scale resources, they are limited to synthesized charts and template-based questions. They do not fully capture the complexity and diversity of real-world charts.

To create more complex QA benchmarks, ChartQA (Masry et al., 2022) mixes 30k human and machine-generated questions; however the images are still limited to line, bar and pie charts. To cover more types, ArXivQA (Li et al., 2024) extracts images with LLM-generated QA pairs from arXiv papers. SciGraphQA (Li and Tajbakhsh, 2023) extract graphs from Comp. Sci. ArXiv papers and use LLMs to create multi-turn question-answering dialogues about them. MMC (Liu et al., 2024) supports diverse tasks and chart types using free-form questions and open-ended answers.

Previous benchmarks rely heavily on chart annotations or table metadata as textual prompts to generate content, allowing models to easily obtain candidate answers while ignoring the charts' visual logic. ChartBench (Xu et al., 2023) includes both annotated and unannotated charts. While ChartX (Xia et al., 2024) covers more chart types, its data and charts are synthesized and limited to ones that can be directly converted into a structural data format, e.g., CSV format. CharXiv (Wang et al., 2024) consists of 2k real-world charts with manually curated questions by human experts and answers validated by hand, panning 8 major subjects published on arXiv. The questions are either descriptive to understand basic chart data or reasoning-based to dig deeper into charts. MultiChartQA (Zhu et al., 2025) is designed to evaluate VLMs' reasoning capabilities across multiple charts. However, the charts are not information-rich and no domain knowledge beyond what is stated in the charts is required to answer questions in these benchmarks.

To cover more image types, MMMU (Yue et al., 2024a,b) collected multimodal questions from college exams, quizzes, and textbooks, covering six disciplines, ranging from visual scenes like photos and paintings to diagrams and tables, testing the perceptual capabilities of VLMs. CURIE (Cui et al., 2025) covers diverse scientific disciplines, but its multimodal tasks are limited to biodiversity georeferencing and protein sequence reconstruction tasks. EMMA (Hao et al., 2025) targets organic multimodal reasoning across mathematics, physics, chemistry, and coding. While these datasets require domain knowledge, it doesn't require a high level of expertise.

Most existing multimodal benchmarks are de-

signed such that reasoning over just the images can result in the answer. But SPIQA (Pramanick et al., 2024) is designed such that questions require simultaneous reasoning over different modalities, including figures, tables, and texts from articles in the computer science domain. Even with the diversity of multimodal benchmarks for scientific literature, there is a dearth of datasets to test how well models can verify claims made in such data.

Scientific Claim Verification Claim verification as the task of establishing the truthfulness of a given claim has gained a lot of attention given ever-increasing amounts of data (Thorne et al., 2018; Kotonya and Toni, 2020b; Wadden et al., 2020). Scientific claim verification requires significant domain knowledge as well as understanding the evidence to reason about the claim. SciFact-Open (Wadden et al., 2022b) expand on previous work (Wadden et al., 2020) to provide a more realistic testbed of claim verification systems. Recent work has also focused on testing how well models can verify claims over tabular data on real-world public health claims and scientific papers (Akhtar et al., 2022; Wang et al., 2021) or charts and images (Akhtar et al., 2024) or a mix of all (Singh et al., 2024). Akhtar et al. (2023, 2024) focus on claim verification over plots and charts. However, the associated plots are often simple and do not adequately test domain knowledge or how to find evidence within larger amounts of data. Our dataset tests domain-specific claim verification abilities over heterogeneous, information-rich figures.

3 Creating MuSciClaims

Verifying whether a claim is supported by scientific evidence requires understanding different parts of the claim, locating and extracting information from multimodal sources, reasoning with it, and finally making a judgment. In scientific articles, such evidence is often presented graphically, in panels of information-rich figures along with a descriptive caption. To test whether models can verify claims over scientific figures, we need claims that are supported, as well as ones that are not. The former can be extracted from papers but manual intervention is required to create the latter. To go beyond standard quantitative benchmarking and better understand model failures, we also need tests that are diagnostic in nature. To this end, we introduce MuSciClaims, a dataset created from scientific articles across physics, chemistry and biology. We use open-access, peer-reviewed articles published in Nature Physics, Journal of the American Chemical Society, and Cell, as the source for claims in physics, chemistry and biology respectively.

3.1 Automatic Extraction

We extract figures and associated claims from the results section of the articles, where key findings and takeaways are described along with supporting evidence, often expressed within heterogeneous figures containing charts, microscopy images, chemical reaction schemes, or other diagrams.

Figures in these articles are quite diverse–they vary in size, resolution, placement, and caption style. We make use of both the HTML and PDF versions of the articles to obtain a uniform organization and representation of all figures. We ensure that only high-resolution images (300ppi) are retained and preprocess the captions to remove irrelevant information such as structural prefixes.

To extract claims associated with the figures, we process the *Results* section text. We use simple regular expressions to identify sentences which either contain explicit references to figures (e.g. "Fig.") or some form of inline references (e.g., "Author et al."). We discard sentences referring to multiple figures or supplementary figures. Full details of the extraction process are provided in Appendix C.

3.2 Systematic Claim Perturbation

The claims extracted from the articles are grounded in the associated figures—i.e., the figures support[3] the claims. To create an effective test bed, we also need claims that are *not supported* and have *contradictory* evidence in the figures. Further, we want to ensure that we test for a variety of reasons that could make a claim unsupported or contradictory with respect to the given evidence. To this end, we use a manual claim perturbation process to ensure meaningful perturbations, and ensure their quality through a second annotation process. Annotators produce free-form contradicting claims over a range of perturbations.

We manually analyzed the original claims to identify the main capabilities needed for checking a claim against the corresponding figures. Based on this, we create four categories of perturbations: (i) *Qualitative Inference*—Directional terms are replaced with their opposites (e.g., "high concentration" to "low concentration"). This tests whether

[3]We rely on the scientific integrity of the published articles and assume that evidence support the asserted claims.

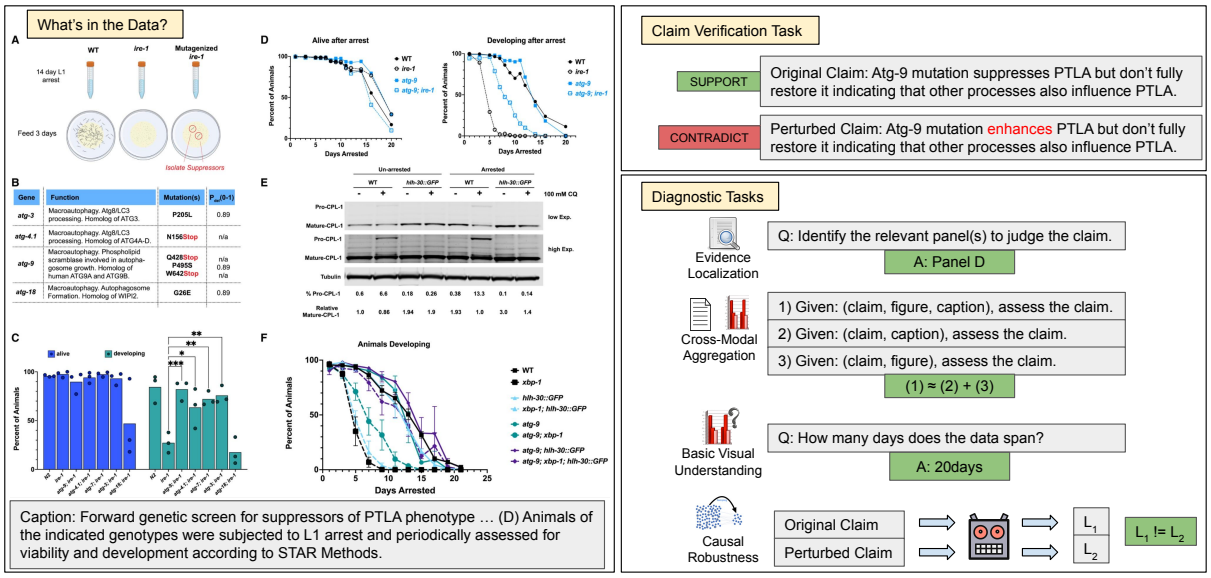


Figure 1: Each data point from MUSCICLAIMS contains a claim, its associated figure and caption and annotations about its relevant panels. Each claim, both original and perturbed, is also labeled with its relationship to the figure (SUPPORT, NEUTRAL, CONTRADICT). It also enables performing diagnostic tests. Models must identify the relevant panel as part of EVIDENCELOCALIZATION and answer a question about the figure for BASICVISUALUNDERSTANDING. To perform CROSS-MODALAGGREGATION, model performance should drop when given either just the figure or the caption. For EPISTEMICSENSITIVITY, model predictions must change across a pair of `(original, perturbed)` claims.

models can check if the asserted qualitative statements are supported via visual relationships between data points in a figure. (ii) *Qualitative Relationship Inference*—Comparisons are edited (e.g., "X is stronger than Y" into "Y is stronger than X") to create the opposite conclusion from a figure. This checks for assessing qualitative relationships between variables via visual inference of similar relationships, (iii) *Quantitative reasoning*—Numerical values, primarily associated with experiment details such as statistical significance or experiment size, are modified to test for reasoning about the key quantities of interest. (iv) *Epistemic Mismatch*—This represents a disconnect between different forms of knowledge. We add perturbations that introduce inconsistencies between an observation that is visually true (i.e., supported by the figure) and its inference (which requires domain knowledge). This tests for the ability to carefully connect visually verified information with the textually asserted effect.

Our perturbations ensure that the modified claim is a contradiction of the supported claim. This means that the figure which supports the original claim will, by extension, not support the modified claim. We verify the quality of the resulting perturbations through a second round of manual annotation. For a subset of data, three annotators were provided a supported claim as well as its perturbation. They are required to judge whether the perturbation contradicts the supported claim. We find that all three annotators agree that all perturbed claims are indeed contradictions of the corresponding supported claims (100% agreement).

3.3 Diagnostics

MUSCICLAIMS is also designed to be a diagnostic dataset to support deeper understanding of model capabilities. We introduce four kinds of diagnostic tests that relate to different aspects of the claim verification problem.

(i) BASICVISUALUNDERSTANDING—For each claim, we identify a data point that is integral to it, and then introduce questions that test for model ability to read or extract it from the figure. (ii) EVIDENCELOCALIZATION—The dataset contains automatically extracted annotations about the panels of a figure which should be perused to judge a claim. Using this, we can test a model's EVIDENCELOCALIZATION ability in the visual modality. (iii) CROSS-MODALAGGREGATION—Often, information (such as that about statistical significance) in the caption is also important to correctly assess a claim. Textual reasoning over the caption

must be combined with visual reasoning over the figure. We test models' multimodal reasoning abilities through CROSS-MODALAGGREGATION. (iv) EPISTEMICSENSITIVITY—Claims often contain an observation from a figure, as well as an inference that also requires domain knowledge to understand. Relationships between observations and inferences are systematically perturbed by annotators as part of §3.2. We collect annotations about whether it is the observation, the inference, or both, that are perturbed. Through EPISTEMICSENSITIVITY, we establish how models change their judgments for such perturbations, indicating their understanding of epistemic relationships within claims. Examples of these diagnostics are provided in Figure 1.

3.4 Dataset Statistics

Through the process described above, we obtain 505 claims that are supported by a figure (SUPPORT), and 505 corresponding perturbed claims in contradiction to a given figure (CONTRADICT). Further, we pair each claim with an unassociated figure from the same paper to obtain data where there is no connection between them (NEUTRAL). Therefore, MUSCICLAIMS contains 1515 data points balanced equally across 3 class labels. Out of this, 918 data points are from biology, 309 from chemistry and 288 from physics. Each data point is also annotated with figure panels most relevant to a claim, a question about the figure and information about perturbation types to support our diagnostic tests[4].

4 Experimental Setup

We benchmark the performance of several state-of-the-art vision-language models (VLMs) on evaluation tasks supported by MUSCICLAIMS.

4.1 Evaluation Tasks

MUSCICLAIMS is designed as a CLAIMVERIFICATION task. Each data point contains a claim, an associated (multi-panel) figure (and caption) and a label (SUPPORT, NEUTRAL, CONTRADICT). Given the figure (and caption) and a claim, models must generate a prediction about whether the claim is supported. We evaluate models on this task using standard metrics of precision, recall and F1 score.

MUSCICLAIMS also supports four *diagnostic* tasks designed to assess a diverse set of capabilities required to effectively verify claims. Performance on these diagnostics highlight limitations of contemporary models, thereby opening up avenues for future research. (1) EVIDENCELOCALIZATION tests whether models can localize to the correct panel(s) in the figure. Given the figure (and caption) and a claim, models must generate the relevant panel names as well as generate a prediction (CLAIMVERIFICATION). We use precision, recall and F1 to measure how well models identify the correct panels. (2) BASICVISUALUNDERSTANDING aims to test whether models can read scientific figures by how models answer a question about the figure. Each claim in MUSCICLAIMS is accompanied by a basic question and its (one-word) answer about the associated figure. We use Exact Match to judge whether a model answer is correct. (3) CROSS-MODALAGGREGATION are experiments designed to analyze how models use the figure and its caption to come up with their judgment. Models need to aggregate information from the figure (visual information) as well as caption (textual information) for claim verification. First, models are given a claim, the associated figure, its caption and required to perform CLAIMVERIFICATION. Then, for the same task, they are prompted to reason over just the figure and just the caption, testing its visual and textual abilities respectively. (4) EPISTEMICSENSITIVITY tests whether models consistently (and correctly) change their prediction across epistemic perturbations of the same claim; they should predict support for the original and contradict for the perturbed claim. Claims often encode epistemic information—observations from the figure and related inferences made with domain knowledge. As part of §3.2, annotators also mark whether they perturb the observation, the inference or both. We perform a sensitivity test for the same across `(original, perturbed)` claim pairs.

4.2 Models

We conduct the aforementioned evaluation on a set of 12 different vision-language models (VLMs): gpt-4o-mini-2024-07-18 (`4o-mini`), gpt-4o-2024-11-20 (`4o`), claude-3-5-sonnet-20241022 (`Sonnet`), o3-2025-04-16 (`o3`), o4-mini-2025-04-16 (`o4-mini`), Phi-4 Multimodal Instruct (`Phi-4`), llava-v1.6-mistral-7b-hf (`Llava-Next`), Llama-3.2-11B Vision Instruct (`Llama-3.2`), Molmo-7B-D (`Molmo`), InternVL3-38B (`InternVL3`), Qwen2.5-VL-32B (`Qwen2.5`) and deepseek-VL2-small (`DeepSeek`). This set represents both open and

[4]We release the data in accordance with the papers' CC BY 4.0 license.

		SUPPORT			NEUTRAL			CONTRADICT			OVERALL		
		P	R	F	P	R	F	P	R	F	P	R	F
4o-mini	D	0.41	0.88	0.56	0.64	0.48	0.55	0.75	0.10	0.17	0.60	0.48	0.43
	R→D	0.43	0.83	0.56	0.64	0.47	0.54	0.62	0.20	0.30	0.56	0.50	0.47
4o	D	0.43	0.93	0.59	0.86	0.46	0.60	0.75	0.23	0.35	0.68	0.54	0.51
	R→D	0.47	0.86	0.61	0.71	0.61	0.65	0.76	0.25	0.38	0.65	0.57	0.55
Sonnet	D	0.52	0.87	0.65	0.83	0.64	0.72	0.79	0.43	0.56	0.71	0.65	0.64
	R→D	0.53	0.89	0.66	0.84	0.64	0.73	0.81	0.47	0.59	**0.73**	0.66	0.66
o3	R→D	**0.67**	0.74	**0.71**	0.69	0.79	0.74	0.82	0.61	**0.70**	**0.73**	**0.72**	**0.72**
o4-mini	R→D	0.62	0.84	**0.71**	0.76	0.76	**0.76**	0.88	0.57	0.69	0.75	**0.72**	**0.72**
Phi-4	D	0.43	0.70	0.53	0.74	0.19	0.30	0.44	0.50	0.47	0.54	0.46	0.43
	R→D	0.36	0.81	0.51	0.81	0.10	0.17	0.58	0.26	0.36	0.58	0.41	0.34
Llava-Next	D	0.37	**0.92**	0.53	0.68	0.33	0.44	**1.00**	0.00	0.01	0.68	0.42	0.33
	R→D	0.38	0.80	0.52	0.54	0.42	0.47	0.61	0.09	0.15	0.51	0.43	0.38
Llama-3.2	D	0.41	0.86	0.56	0.68	0.42	0.52	0.60	0.17	0.27	0.57	0.49	0.45
	R→D	0.37	0.93	0.53	0.72	0.15	0.25	0.61	0.17	0.27	0.56	0.42	0.35
Molmo	D	0.41	0.91	0.57	0.77	0.29	0.42	0.54	0.22	0.32	0.57	0.47	0.43
	R→D	0.39	0.75	0.51	0.56	0.28	0.37	0.43	0.23	0.30	0.46	0.42	0.39
InternVL3	D	0.62	0.79	0.70	**0.83**	0.68	0.75	0.70	**0.64**	0.67	0.72	0.70	0.70
	R→D	0.45	0.91	0.60	**0.83**	0.49	0.61	0.80	0.30	0.44	0.69	0.57	0.55
Qwen2.5	D	0.55	0.80	0.65	0.70	**0.80**	0.75	0.85	0.34	0.48	0.70	0.65	0.63
	R→D	0.42	0.91	0.57	0.79	0.41	0.54	0.85	0.25	0.39	0.68	0.53	0.50
DeepSeek	D	0.55	0.43	0.48	0.50	0.67	0.58	0.44	0.40	0.42	0.50	0.67	0.58
	R→D	0.41	0.65	0.51	0.51	0.51	0.51	0.51	0.21	0.30	0.48	0.46	0.44

Table 1: Model performance on the claim verification task of MUSCICLAIMS when prompted to simply generate the decision (D), and when asked to reason and then generating the decision (R→D). InternVL3 achieves best performance when prompted to just give the answer, while o3 and o4-mini are the best overall, using their inbuilt reasoning capabilities. Closed-source models are slightly better with reasoning whereas open-source models do worse in most cases, represent a significant gap in their reasoning capabilities.

closed-sourced models of differing capabilities for a comprehensive evaluation of MUSCICLAIMS. We evaluate models primarily in two zero-shot settings: (i) generating only a judgment (D), and (ii) reasoning about the claim before judging it (R→D). More details are in Appendix A, G and D.

5 Results

Table 1 presents the performance of all the models in different settings for the multimodal scientific claim verification task. We present per-class (SUPPORT, NEUTRAL and CONTRADICT) precision, recall and F1 score as well as macro average metrics on the class balanced MUSCICLAIMS[5]. We make two main observations.

Most VLMs perform poorly on MUSCICLAIMS. We observe that most models perform poorly on the task (D rows in Table 1), with overall F1 scores only ranging from ∼0.3-0.5. Only two models (out of ten) stand out: Sonnet (0.66 F1) and InternVL3 (0.70 F1). Models attain high recall and low precision on SUPPORT. In contrast for the NEUTRAL and CONTRADICT claims, models have low recall and high precision. This implies two findings: First, models have a strong bias towards recognizing most of the claims as supported. Second, the models can reliably identify some of the NEUTRAL and CONTRADICT claims. Our manual analysis shows that models only identify the most obviously wrong claims as the NEUTRAL and CONTRADICT claims. This can be done reliably but they struggle on ones that are more difficult, which require careful reasoning. These suggest challenges for claim verification methods.

Reasoning before judging helps models slightly. The R→D rows in Table 1 show results where models, given the figure and caption, first perform step-by-step reasoning on the claim and then generate

[5]Results for each domain are presented in Table 8, 9 and 10

their decision on the category of the claim. Results show that reasoning leads to improvements (∼0.02-0.04) for closed-source models and `Llava-Next`, but the gains are rather small. `o3` and `o4-mini`, models trained to analyze and do reasoning over images, achieve the highest performance (0.72 F1).

There is a notable drop in performance for open-source models (∼0.04-0.16) indicating a weakness in CoT abilities of open-source models for claim verification. We hypothesize that this is due to the limitations of instruction tuning in vision-language modeling where models are mainly finetuned to describe or analyze images, not reasoning chains.

Table 7 presents model performance for different domains in MUSCICLAIMS. On average, models are worst at verifying PHYSICS claims and best at judging CHEMISTRY claims. However, the highest performance is achieved on BIOLOGY claims.

6 Diagnostics Results

We use our diagnostic tests to better understand the failure modes of `4o-mini`, `4o`, `Sonnet` and `InternVL3`. Going forward, we run these diagnostic tests (§3.3) on the BIOLOGY subset of MUSCICLAIMS and discuss them.

		P	R	F
`4o-mini`	R→D	0.57	0.50	0.46
	I→R→D	0.59	0.45	0.40
`4o`	R→D	0.69	0.59	0.56
	I→R→D	0.73	0.51	0.47
`Sonnet`	R→D	0.78	**0.70**	**0.70**
	I→R→D	**0.79**	0.69	**0.70**
`InternVL3`	R→D	0.75	0.59	0.58
	I→R→D	0.75	0.59	0.58

Table 2: Model performance on claim verification worsens when also prompted to localize to the relevant panels (I→R→D) as compared to reasoning over the entire figure and assessing a claim (R→D).

VLMs localize poorly to relevant information. Finding the most relevant panels of figures is important to assess claims from information-rich figures. Table 2 shows how well models perform when prompted to first identify the associated panels, reason over them and make a decision (I→R→D), thereby testing EVIDENCELOCALIZATION. Model performance deteriorates when localizing before reasoning (I→R→D) as compared to reasoning over the entire figure (R→D). We also explicitly test their ability to locate relevant panels. Table 3 presents precision, recall and F1 to measure how well models can localize to the correct visual evidence. Their low precision and high recall indicates they do identify the relevant panel(s), but also deem a lot of irrelevant panels to be important. Clearly, evidence localization is difficult for models.

	P	R	F
`4o-mini`	0.37	0.77	0.50
`4o`	0.53	0.70	0.61
`Sonnet`	**0.62**	**0.80**	**0.70**
`InternVL3`	0.46	0.68	0.55

Table 3: EVIDENCELOCALIZATION—We use precision, recall and F1 score to characterize how well models can localize to relevant panels. Low precision indicates that they are bad at identifying only the correct panels.

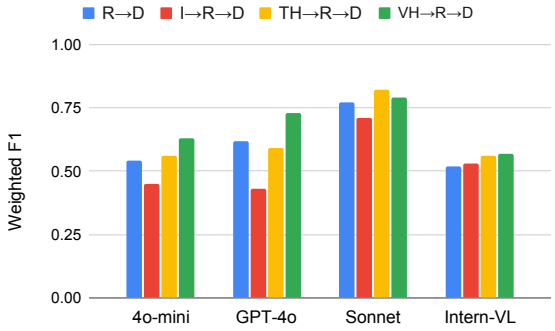


Figure 2: Model performance on CLAIMVERIFICATION when prompted for (or provided) localization. Providing models with hints about the relevant panels of figures improves their claim verification. Textual hints (TH) guide models to focus on the correct part of the full figure, showing higher performance than R→D and I→R→D. Models using visual hints (VH; relevant panel as visual input instead of full figure) perform even better. This indicates that localizing to the relevant knowledge has the potential to improve models.

Better localization can improve performance. We perform a series of experiments to establish how well models can perform if they have correct localization information. First, we provide models gold information about which panels are associated with the claim as a *textual hint* (TH→R→D). Next, for each claim, instead of the full figure, we only provide the relevant panel to the model as a *visual hint* (VH→R→D), instead of the full figure. These

experiments are performed over a randomly sampled subset (n=101) of class-balanced data points.

Figure 2 compares the performance of models with and without these hints. As stated earlier, models are better at reasoning over the full figure (R→D) rather than over panels it has identified as relevant (I→R→D). However, when given the relevant panels as a textual hint (TH→R→D), they fare much better. They improve even further when only given the relevant panel(s) of the figure (VH→R→D) as input, thus removing panel localization errors. The poor localization performance coupled with the gains seen with localization hints suggest that improving the localization abilities of models is valuable. But even with perfect localization (i.e., through hints here), there is significant room for improvement, indicating challenges in other aspects of multimodal reasoning.

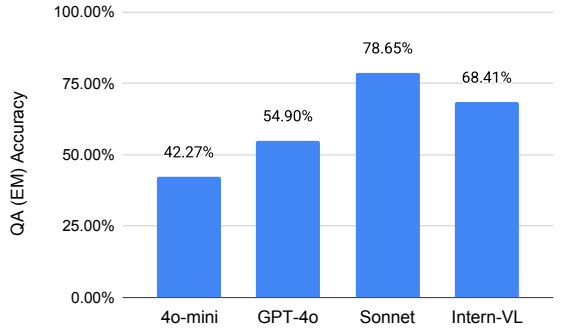


Figure 3: Model performance on BASICVISUALUNDERSTANDING. Models fail to answer basic questions about components of figures associated with claims.

Models need to improve on basic visual reading. Each claim in MUSCICLAIMS is also accompanied by a question about the figure that is relevant for the verification process. These questions test basic visual reading abilities (e.g., "How many days does the data span?" in Figure 1) and do not require complex reasoning. Figure 3 shows that most models perform poorly on such questions. `Sonnet` performs the best, correctly answering ∼78% of the questions. The moderate performance indicates a gap in models' visual comprehension capabilities when it comes to scientific figures.

Models struggle with cross-modal reasoning. Table 4 compares model performance when provided both the figure and its caption, just the caption and just the figure. Models must reason over information in both modalities in order to best assess a claim since information is found in both the figure (visual) as well as its caption (textual). However, we note that models' performance doesn't improve substantially over its performance when using just one modality. This indicates that they might not be effectively combining the complementary information present in both modalities.

		F+C	C	F
`4o-mini`	D	0.42	0.46	0.38
	R→D	0.46	0.50	0.45
`4o`	D	0.52	0.50	0.45
	R→D	0.56	0.44	0.51
`Sonnet`	D	0.68	0.58	0.60
	R→D	0.70	0.51	0.64
`InternVL3`	D	0.74	0.64	0.68
	R→D	0.58	0.47	0.47

Table 4: Models achieve a large chunk of their performance using information from just one modality even though information from both modalities is needed to judge claims. F+C indicates when both the figure and the caption is provided, C indicates when only the caption (textual) is provided, and F indicates when only the figure (visual) is provided to the model.

		Obs	Inf	Both	None
4o-mini	D	13%	13%	0%	12%
	R→D	20%	28%	0%	22%
4o	D	13%	30%	0%	23%
	R→D	7%	26%	0%	27%
Sonnet	D	47%	50%	0%	45%
	R→D	47%	54%	0%	52%
InternVL	D	67%	72%	50%	65%
	R→D	60%	39%	0%	35%

Table 5: Model sensitivity—changing their prediction about a claim for different types of perturbation.

Models can't handle epistemic mismatches. Claims often encode epistemic relationships which can be systematically perturbed to test the sensitivity of contemporary models (Verma et al., 2023). We calculate sensitivity as the percentage of times models change predictions across the supported and refuted version of the same claim. Table 5 shows the sensitivity of models by perturbation type. Models are not sensitive enough to understand nuances in epistemic relationships, being the

	# Examples	SUPPORT			NEUTRAL			CONTRADICT			OVERALL		
		P	R	F	P	R	F	P	R	F	P	R	F
4o-mini	$k=0$	0.40	0.90	0.56	0.66	0.42	0.51	0.74	0.10	0.18	0.6	0.47	0.42
	$k=1$	0.41	0.90	0.56	0.81	0.12	0.22	0.73	0.48	0.58	0.65	0.50	0.45
	$k=3$	0.41	0.92	0.57	0.71	0.46	0.56	0.8	0.11	0.19	0.64	0.50	0.44
	$k=5$	0.41	0.91	0.56	0.71	0.49	0.58	0.86	0.08	0.15	0.66	0.49	0.43
GPT-4o	$k=0$	0.43	0.96	0.59	0.93	0.44	0.60	0.84	0.23	0.36	0.73	0.54	0.52
	$k=1$	0.48	0.92	0.63	0.90	0.53	0.66	0.77	0.38	0.51	0.72	0.61	0.60
	$k=3$	0.46	0.95	0.62	0.90	0.53	0.67	0.83	0.28	0.42	0.73	0.59	0.57
	$k=5$	0.45	0.95	0.62	0.90	0.51	0.65	0.86	0.28	0.42	0.74	0.58	0.56
Sonnet	$k=0$	0.53	0.92	0.67	0.91	0.63	0.74	0.82	0.49	0.61	0.76	0.68	0.68
	$k=1$	0.64	0.73	0.68	0.74	0.74	0.74	0.76	0.66	0.71	0.71	0.71	0.71
	$k=3$	0.62	0.75	0.68	0.75	0.66	0.70	0.73	0.65	0.69	0.70	0.69	0.69
	$k=5$	0.60	0.83	0.69	0.83	0.69	0.76	0.77	0.60	0.67	0.73	0.71	0.71

Table 6: Model (D) performance on BIOLOGY claims in MUSCICLAIMS. k denotes the number of few-shot examples provided as part of the prompt.

worst when both the observation and inference is modified. Analyzing differences in models' confidences for predictions may provide more insight (Marcé and Poliak, 2022).

Few-shot examples help a little. We experiment with few-shot prompting (or in-context learning) (Brown et al., 2020; Wei et al., 2022) on BIOLOGY claims for the decision-only (D) experiments using a subset of closed-source models. Using the methodology described in section 3, we create 45 claims from Cell papers [6]. The few-shot examples (k) are selected randomly from these created claims.

Table 6 presents model performance when prompted to just produce a decision (D). We find that performance improves for all models when they are provided any number of in-context examples. However, the benefits go down as the number of examples goes up.

7 Conclusion

Assessing whether claims are supported requires understanding the methods and data presented in associated figures. One must find the correct piece of information in the figure and then combine it with the caption. This paper introduces MUSCICLAIMS, a new diagnostic dataset to evaluate the claim verification capabilities of VLMs. We find that most VLMs are poor at this task out-of-the-box, and chain-of-thought only helps slightly. Particularly, they are significantly worse at understanding that given evidence contradicts (or is not related to) the claim. EVIDENCELOCALIZATION shows that models are bad at identifying the right panel of data, a critical flaw in their claim verification capabilities. CROSS-MODALAGGREGATION indicates that models do not effectively use both visual and textual information for their judgments. Diagnostics also reveal that they do not understand some obvious characteristics of the associated figures. Our results establish the current abilities of VLMs for claim verification over heterogeneous, information-rich scientific figures, and our diagnostics highlight specific avenues of research to improve them.

Limitations

We benchmark a reasonably diverse set of VLMs. However, we acknowledge that we can try more models across a spectrum of architectures, training paradigms and sizes. Due to the current fast-paced landscape of VLM development, we will continue to evaluate more VLMs on MUSCICLAIMS.

Due to the difficulty of creating data that VLMs have not already seen (published after their cutoff dates), we are unable to train models for this task. We perform few-shot experiments with closed-source models (`4o-mini`, `4o` and `Sonnet`) but we leave further exploration of different methods of example selection to future work.

We formulate the task of multimodal scientific claim verification. But our dataset is limited to using captions as the textual part of the input to models. While these captions are descriptive, models might benefit from using extra context, such as that extracted from the *Methods* sections of papers. Further, we limit the extraction of claims only to the *Results* sections, even though claims may occur

[6] These papers are not part of the MUSCICLAIMS evaluation set.

in other sections too.

In this work, we perform some qualitative evaluation of weaknesses in the reasoning produced by VLMs. However, we are unable to do so at scale. Such evaluation requires experts with incredibly specific domain expertise. Even a graduate student (PhD level) or faculty cannot verify reasoning for all domains covered in Nature Physics, Journal of the American Chemical Society and Cell. For instance, an expert in ecology cannot easily judge the reasoning about claims in cellular biology. We will explore how to conduct better evaluations as part of future work.

Our work only investigates English-language documents and this limits the generalizability of our findings to other languages, although most scientific articles are disseminated in English.

Due to high cost of the recently released `o3` and `o4-mini` models, we are unable to analyze it across the full spectrum of our diagnostics. For consistency, we analyze `Sonnet` and `InternVL3` since they have similar performance on MuSciClaims.

Ethical Considerations and Risks

Prior work has shown that VLMs exhibit various types of bias. While they do not generate free-form language for our binary prediction task, it is possible, though highly unlikely, that biases explicitly come up in the explanations. Deploying such unreliable models into critical infrastructure and relying on them for decisions can cause harm to users.

Acknowledgments

This material is based on research that is in part supported by the DARPA for the SciFy program under agreement number HR00112520301. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of DARPA or the U.S. Government. The authors would also like to thank Aakanksha Rajiv Kapoor for her help with understanding the structure and content of papers published in the Cell journal and for help with the qualitative error analysis of model outputs.

References

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. ChartCheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937, Bangkok, Thailand. Association for Computational Linguistics.

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

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, and 1 others. 2025. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. *arXiv preprint arXiv:2503.13517*.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can MLLMs reason in multimodality? EMMA: An enhanced multimodal reasoning benchmark. In *Forty-second International Conference on Machine Learning*.

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

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

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

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

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *NAACL-HLT*.

Sanjana Marcé and Adam Poliak. 2022. On gender biases in offensive language classification models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 174–183, Seattle, Washington. 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.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *arXiv preprint arXiv:2407.09413*.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298.

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.

Dhruv Verma, Yash Kumar Lal, Shreyashee Sinha, Benjamin Van Durme, and Adam Poliak. 2023. Evaluating paraphrastic robustness in textual entailment models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 880–892, Toronto, Canada. 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.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022b. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, and 1 others. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2025. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11341–11359, Albuquerque, New Mexico. Association for Computational Linguistics.

A Benchmark Models

We provide details of each model we evaluate on MUSCICLAIMS.

gpt-4o-2024-11-20 accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. It is especially better at vision and audio understanding compared to existing models.

gpt-4o-mini-2024-07-18 has a context window of 128K tokens, supports up to 16K output tokens per request. It surpasses other small models released to that date on academic benchmarks across both textual intelligence and multimodal reasoning, and supports the same range of languages as 4o.

claude-3-5-sonnet-20241022 sets new industry benchmarks for graduate-level reasoning (GPQA), undergraduate-level knowledge (MMLU), and coding proficiency (HumanEval). It shows marked improvement in grasping nuance, humor, and complex instructions, and is exceptional at writing high-quality content with a natural, relatable tone.

o3-2025-04-16 excels at solving complex math, coding, and scientific challenges while demonstrating strong visual perception and analysis. It uses tools in its chains of thought to augment its capabilities; for example, cropping or transforming images, searching the web, or using Python to analyze data during the thought process.

o4-mini-2025-04-16 is a smaller model optimized for fast, cost-efficient reasoning—it achieves remarkable performance for its size and cost, particularly in math, coding, and visual tasks. It is the best-performing benchmarked model on AIME 2024 and 2025. It performs especially strongly at visual tasks like analyzing images, charts, and graphics.

Phi-4-multimodal-instruct is a 5.6 billion parameter multimodal modal that combines image, textual and audio modalaties into a single small language model via LoRA adapters and modality-specific routers that make multiple inference modes possible without interference. The model has been extensively instruction tuned on a combination of synthetic and web data.

llava-v1.6-mistral-7b-hf is a 7.6 billion parameter vision language model that is part of the Llava-Next regimen and built on top of the Llava architecture. It has a pretrained vision encoder and Mistral-7B as the language modeling backbone. It has been instruction tuned on over a million data points coming from a combination of high-quality user instruct data and multimodal document/chart.

Llama-3.2-11B-Vision-Instruct is the 11B version of the Llama 3.2-Vision set of multimodal LLMs which have been instruction tuned for image reasoning. It is built on top of the pretrained Llama 3.1 text only LLM by combining a seperately trained vision adapter module. Using a combination of supervised fine-tuning and reinforcement learning from human feedback, the model has been optimized to do a variety of vision tasks like image recognition, reasoning, captioning, and question answering on images.

Molmo-7B-D-0924 is a 7 billion parameter open-source vision-language model. It is developed upon the Qwen2-7B language model with OpenAI CLIP as the vision adapter. The model has been trained on PiXMo, a dataset containing 1 million high quality curated (image,text) tuples.

InternVL3-38B is a 38 billion parameter open-source vision language model. It has been built based upon the following components: variable visual position encoding which handles longer multimodal context; native multimodal pre-training that combines language pre-training and multimodal post-training in a single pipeline; mixed preference optimization to align the model response distribution with the ground-truth distribution; and test-time scaling using VisualPRM-8B as a critic model for Best-of-N evaluation.

Qwen2.5-VL-32B-Instruct is a 32 billion parameter vision language model. It is created on top of the Qwen-2.5 7 billion language model by following the ViT architecture. It has been extensively instruction tuned on (image,text) tuples to so that the model understands all things visual, is agentic, can comprehend long videos and events, can do visual localization, and generate structured outputs.

deepseek-vl2-small is a 16 billion parameters mixture-of-experts vision language model. It has shown been to demonstrate enhanced performance across multiple tasks like visual question

answering, optical character recognition, document/table/chart understanding, and visual grounding. It improves upon its predecessor, DeepSeek-VL, by using an improved high-resoultion vision encoder for better visual comprehension and an optimized language model backbone for training and test time efficiency. It is trained on a data that boosts performance and gives new capabilities to the model such as precise visual grounding.

B Human Annotation Details

The design and instructions for the different applications through which annotations are collected can be found in Figure 4, 5 and 6. Our annotators are graduate students who are experts at reading figures but have limited domain knowledge. To alleviate the skill issue, we ask them to avoid perturbations that require more domain knowledge than they possess. They were not paid for annotations, and were informed of how the annotations would be used.

C Automatic Extraction Details

To construct a high-quality multimodal benchmark for scientific claim verification, we developed an automated pipeline for extracting textual claims and their associated visual elements from research articles. Our approach operates over full-text HTML and PDF documents sourced from Nature Physics (`https://www.nature.com/nphys/`), the Journal of the American Chemical Society (`https://pubs.acs.org/journal/jacsat`) and Cell (`https://www.cell.com`), leading impact factor venues in their respective fields. It creates reliable mapping between complex scientific assertions and their supporting visual evidence with minimal manual supervision. During dataset collection, no personal identifying info (PII) was collected. None of the collected data contained any offensive content.

C.1 Automatic Figure Extraction

We extract figures and their corresponding captions from the structured HTML versions of articles sourced from these journals. Each article contains embedded figure blocks that follow consistent filename conventions and DOM structures, allowing for reliable identification and extraction. Figures are mapped to canonical identifiers (e.g., `figure_1`, `figure_2`, etc.) to ensure consistency across the dataset.

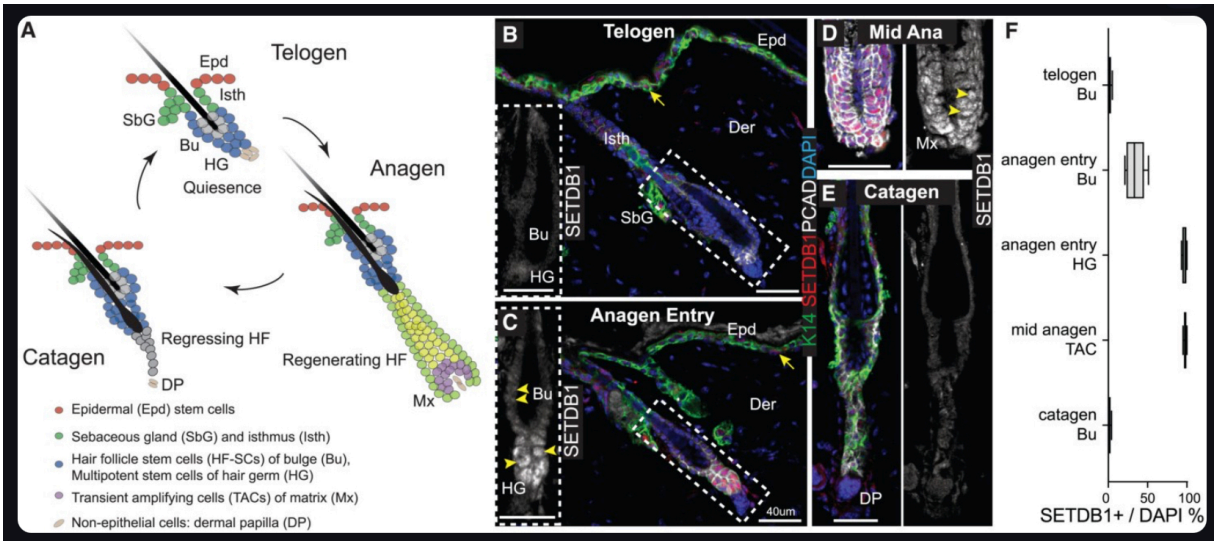
Captions are extracted from the `<figcaption>` elements associated with each figure and typically consist of a short title followed by descriptive text. We concatenate these segments, remove structural prefixes and apply light normalization to clean residual markup or formatting noise. Only high-resolution main-text figures are retained, while supplementary or non-standard assets are excluded. This approach yields a clean, structured mapping between each visual element and its corresponding caption, enabling precise alignment with textual claims during the dataset construction process.

C.2 Automatic Claim Extraction

Scientific claims are typically concentrated in the *Results* section, where authors present novel findings grounded in empirical data, often accompanied by figures such as charts, microscopy images, or diagrams. In contrast, other sections such as *Introduction* or *Discussion* tend to be more speculative, summarizing prior work or offering high-level interpretations. To ensure that extracted statements are factual, visually grounded, and suitable for verification, we restrict claim extraction to the *Results* section.

We process article PDFs using a layout-aware parser to identify the *Results* section and extract its contents. Section headers such as "Results" and "Discussion" are detected using regex patterns robust to formatting variations and numbering conventions. The extracted text is segmented into candidate sentences using a customized version of the NLTK Punkt tokenizer, adapted for scientific prose by accounting for common abbreviations (e.g., "Fig.", "et al.") and inline structures such as references and equations.

Candidate sentences are filtered using a series of quality criteria to ensure that only concise, visually grounded claims are retained. Specifically, each sentence must (i) contain an explicit reference to a main-text figure (e.g., "(Figure 2A)"), (ii) be between 40 and 800 characters in length, (iii) include at least 8 words, and (iv) not match known patterns associated with citations, table fragments, or supplementary material. To maintain clarity and reduce ambiguity during alignment, we retain only single-sentence claims that refer to a single primary figure (i.e., claims with multiple distinct figure references are excluded). Additionally, we restrict figure selection to images smaller than 5MB to ensure compatibility with downstream modeling. This process yields a clean set of scientific claims,



Claim Figure

Caption: Activated multipotent stem cells and TACs in the skin are marked by SETDB1 (A) Schematics of the murine hair cycle, including resting stage "telogen," regeneration stage "anagen," and regression stage "catagen." (B–E) Immunofluorescence (IF) for SETDB1 (red), KERATIN 14 (K14 labeling all skin epithelium, green), P-CADHERIN (PCAD enriched in hair follicle, gray), and DAPI (blue) throughout hair cycle. (B and C) insets and (D and E) single channel images showing SETDB1 (gray) prominent nuclear staining in activated HG (C) and TACs (D), low signal in activated bulge HF-SCs (C), and none in quiescent bulge (B) or during regression (E). Yellow arrows and arrowheads denote SETDB1 nuclear signals in the epidermis (moderate signal) and hair follicles, respectively. (F) Quantification of SETDB1+ cells among DAPI within each anatomic site of the indicated hair cycle stage. Bu, bulge; HG, hair germ; TAC, transient amplifying cell. See

Figure 4: Instructions and UI of the application used to collect perturbations of claims from manual annotators.

each grounded in a single visual source and suitable for fine-grained multimodal verification and localization tasks.

D Model Setup

Python was the main scripting language for data collection and experimentation. For experiments using closed source models, we used OpenAI[7] and Anthropic[8] APIs. The total cost for OpenAI was $\sim$ 400 USD and $\sim$ 160 USD for Claude. The open-source experiments were conducted on 4 A6000 GPUs, each having 48 GB. The total GPU hours for all the experiments was $\sim$40. The models were downloaded from Huggingface and hosted for inference using Huggingface transformers module and vLLM. We use GitHub Co-Pilot to help with writing code but verify it manually before running any experiments.

[7] https://openai.com/api/pricing/

[8] https://://www.anthropic.com/pricing

E Additional Results

E.1 Model Performance by Domain

Table 7 presents model performance for different domains in MUSCICLAIMS. On average, we note that models are worst at verifying PHYSICS claims and best at judging CHEMISTRY claims. However, the highest performance is achieved by `o3` and `o4-mini` on BIOLOGY claims. We also present per-label metrics for model performance in each domain in Table 8, Table 9 and Table 10.

E.2 MUSCICLAIMS task as a two-class problem

Table 11 presents the results when the main claim verification task of MUSCICLAIMS is converted from a three-class (SUPPORT, CONTRADICT, NEUTRAL) problem to a two-class problem by merging CONTRADICT and NEUTRAL classes to NON-SUPPORT class. From the table, we observe that overall F1-scores do vary from the three-class F1-

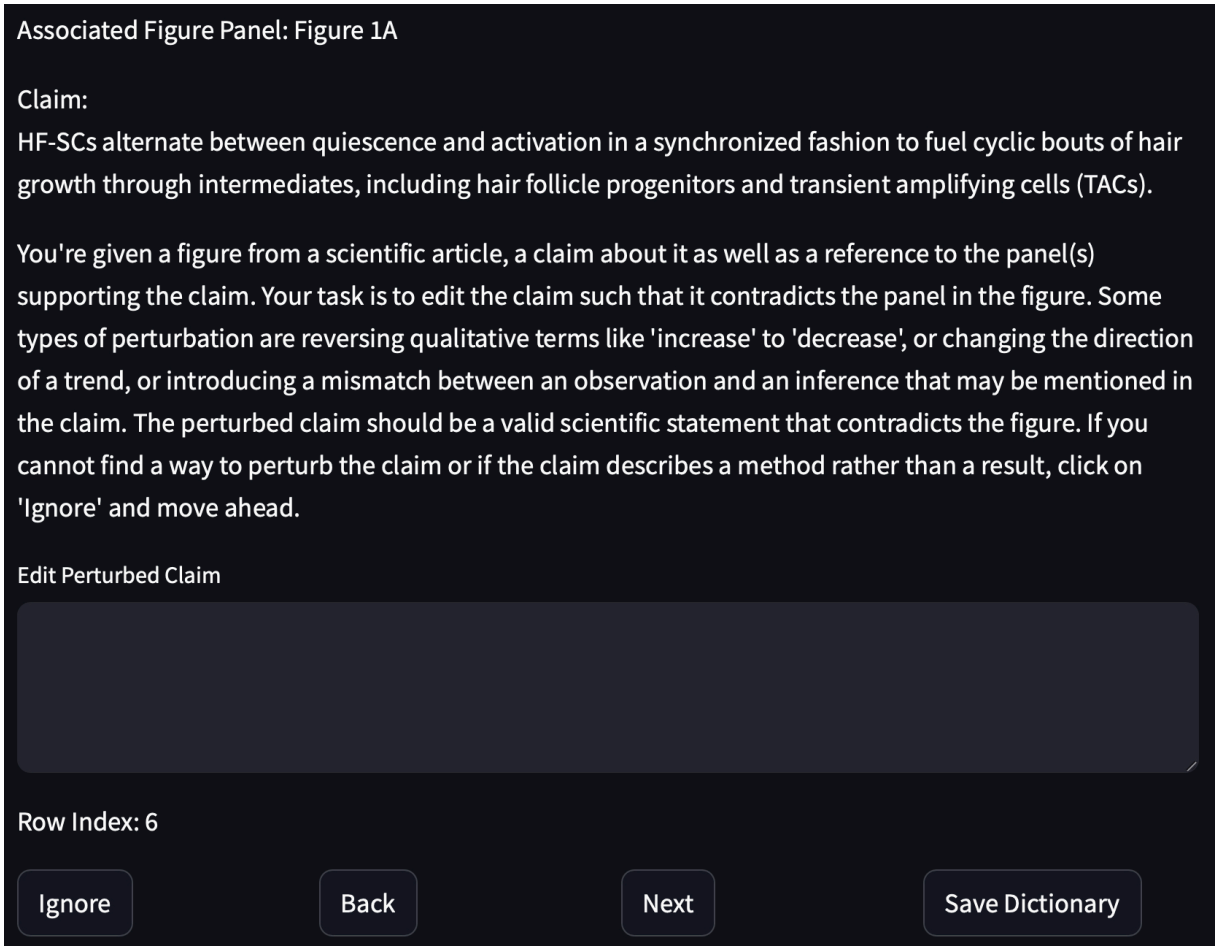


Figure 5: Instructions and UI of the application used to collect perturbations of claims from manual annotators (contd.).

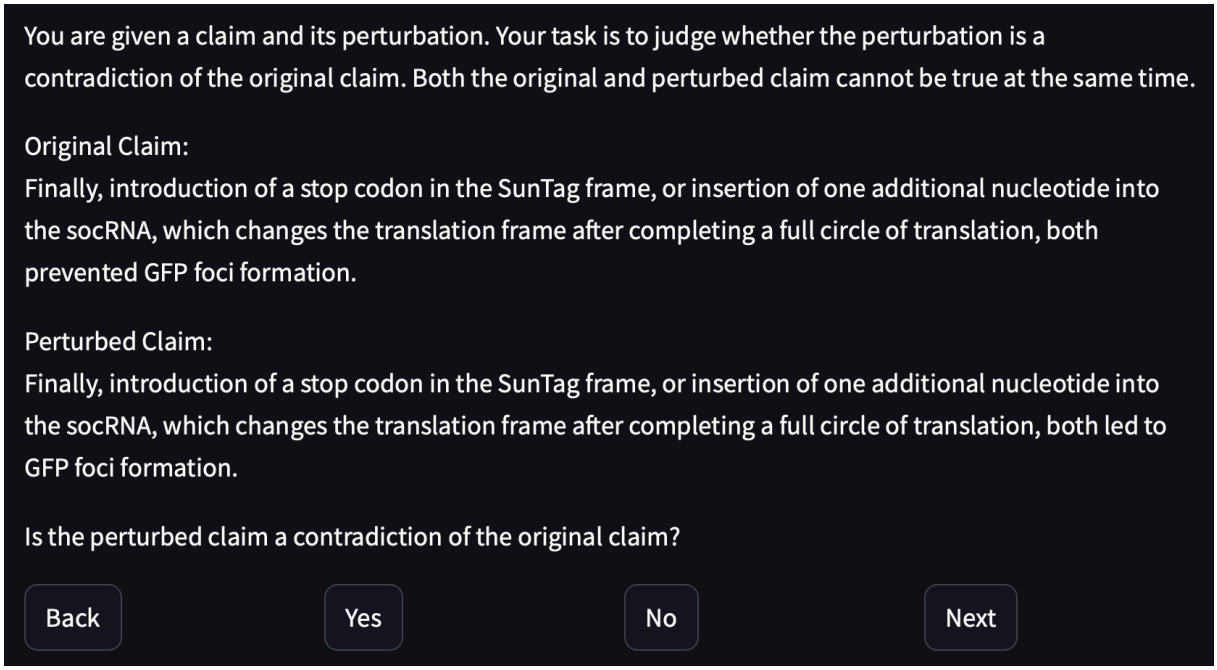


Figure 6: Instructions and UI designed to collect a second round of manual annotation to verify that the perturbed claims are contradictions of the supported claims through Figure 4 and Figure 5. We ask three annotators to do this task and find that they full agree for all the perturbations.

		Physics			Chemistry			Biology			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
4o-mini	D	0.56	0.45	0.38	0.65	0.55	0.49	0.60	0.47	0.42	0.60	0.48	0.43
	R→D	0.52	0.47	0.43	0.57	0.54	0.51	0.57	0.50	0.46	0.56	0.50	0.47
4o	D	0.53	0.47	0.43	0.69	0.59	0.57	0.73	0.54	0.52	0.68	0.54	0.51
	R→D	0.55	0.50	0.48	0.64	0.60	0.57	0.69	0.59	0.56	0.65	0.57	0.55
Sonnet	D	0.62	0.53	0.50	0.68	0.62	0.61	0.76	0.68	0.68	0.71	0.65	0.64
	R→D	0.62	0.53	0.51	0.72	0.66	0.66	0.78	0.70	0.70	0.73	0.66	0.66
o3	R→D	0.58	0.55	0.54	0.72	0.71	0.71	0.79	0.77	0.77	0.73	0.72	0.72
o4-mini	R→D	0.64	0.59	0.57	0.73	0.72	0.71	0.81	0.77	0.77	0.75	0.72	0.72
Phi-4	D	0.43	0.40	0.38	0.60	0.48	0.47	0.63	0.47	0.43	0.54	0.46	0.43
	R→D	0.62	0.37	0.27	0.52	0.39	0.32	0.60	0.42	0.37	0.58	0.41	0.34
Llava-Next	D	0.28	0.39	0.30	0.36	0.46	0.37	0.74	0.42	0.32	0.68	0.42	0.33
	R→D	0.52	0.39	0.34	0.55	0.45	0.40	0.50	0.44	0.38	0.51	0.43	0.38
Llama-3.2	D	0.51	0.45	0.41	0.56	0.48	0.44	0.60	0.50	0.47	0.57	0.49	0.45
	R→D	0.53	0.41	0.35	0.52	0.41	0.34	0.60	0.43	0.36	0.56	0.42	0.35
Molmo	D	0.46	0.41	0.37	0.55	0.47	0.42	0.62	0.50	0.46	0.57	0.47	0.43
	R→D	0.43	0.40	0.38	0.44	0.41	0.37	0.47	0.43	0.41	0.46	0.42	0.39
InternVL3	D	0.59	0.58	0.57	0.73	0.72	0.72	0.77	0.74	0.74	0.72	0.70	0.70
	R→D	0.56	0.49	0.46	0.69	0.57	0.56	0.75	0.59	0.58	0.69	0.57	0.55
Qwen2.5	D	0.68	0.58	0.54	0.70	0.65	0.62	0.72	0.66	0.65	0.70	0.65	0.63
	R→D	0.63	0.47	0.43	0.70	0.53	0.51	0.71	0.54	0.52	0.68	0.53	0.50
DeepSeek	D	0.47	0.47	0.46	0.52	0.50	0.49	0.51	0.51	0.50	0.50	0.67	0.58
	R→D	0.42	0.42	0.39	0.46	0.46	0.42	0.50	0.47	0.46	0.48	0.46	0.44

Table 7: Model performance on the claim verification task of MuSciClaims by the scientific domain of the claims when prompted to simply generate the decision (D), and when asked to reason and then generating the decision (R→D).

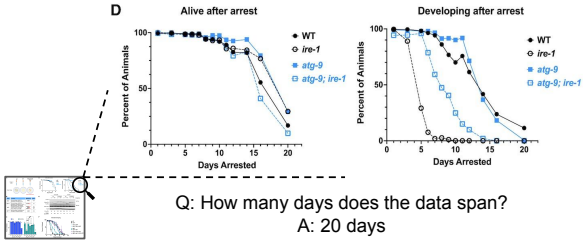


Figure 7: Each claim is accompanied by a diagnostic question that tests whether models can read the relevant panel of the claim's associated figure.

scores, highlighting that MuSciClaims is hard to solve even on a simplified problem setting. We see higher precision and recall values for Non-Support compared Contradict and Neutral metrics in three-class problem. This shows that while models can do coarse-grained classification of wrong or irrelevant claims, in context of the figure and cpation, but struggle when doing fine-grained classification.

E.3 Panel Complexity

Table 12 shows the results of different models when doing inference on single panel images from MuSciClaims compared to multi-panel images. Multi-panel images represent claim verification tasks from MuSciClaims of higher complexity since models have to reason on the correct panel and filter out distractor panels. However, results show models doing better on average for multi-panel setting compared to single-panel setting. This might be because multi-panel provides more visual context for model to do the task.

F Qualitative Error Analysis

In addition to the diagnostics, we perform qualitative error analysis on a random sample of 100 errors in o3 reasoning (R→D) on MuSciClaims. We categorize the errors into the following categories:

1. Domain Expertise (27%) - Models lack domain expertise, knowledge of related work and common practices of representing data specific to

		Support			Neutral			Contradict			Overall		
		P	R	F	P	R	F	P	R	F	P	R	F
4o-mini	D	0.40	0.81	0.54	0.52	0.47	0.49	0.75	0.06	0.12	0.56	0.45	0.38
	R→D	0.41	0.79	0.54	0.56	0.46	0.50	0.58	0.15	0.23	0.52	0.47	0.43
4o	D	0.40	0.83	0.54	0.63	0.38	0.47	0.56	0.19	0.28	0.53	0.47	0.43
	R→D	0.44	0.71	0.54	0.54	0.57	0.56	0.66	0.22	0.33	0.55	0.50	0.48
Sonnet	D	0.44	0.77	0.56	0.60	0.58	0.59	0.81	0.23	0.36	0.62	0.53	0.50
	R→D	0.45	0.79	0.57	0.60	0.54	0.57	0.81	0.26	0.39	0.62	0.53	0.51
o3	R→D	**0.59**	0.52	0.55	0.50	0.79	0.62	0.63	**0.34**	0.45	0.58	0.55	0.54
o4-mini	R→D	0.53	0.66	0.59	0.56	0.76	0.65	0.82	**0.34**	**0.49**	0.64	**0.59**	**0.57**
Phi-4	D	0.37	0.70	0.49	0.43	0.24	0.31	0.48	0.27	0.35	0.43	0.40	0.38
	R→D	0.35	**0.90**	0.50	**1.00**	0.02	0.04	0.51	0.19	0.27	0.62	0.37	0.27
Llava-Next	D	0.36	0.78	0.49	0.47	0.39	0.42	0.00	0.00	0.00	0.28	0.39	0.30
	R→D	0.37	0.74	0.49	0.40	0.35	0.38	0.80	0.08	0.15	0.52	0.39	0.34
Llama-3.2	D	0.40	0.80	0.53	0.54	0.41	0.46	0.59	0.14	0.22	0.51	0.45	0.41
	R→D	0.36	**0.90**	0.52	0.59	0.18	0.27	0.65	0.16	0.25	0.53	0.41	0.35
Molmo	D	0.38	0.83	0.52	0.51	0.22	0.31	0.50	0.19	0.27	0.46	0.41	0.37
	R→D	0.37	0.72	0.49	0.50	0.23	0.31	0.43	0.26	0.32	0.43	0.40	0.38
InternVL3	D	0.55	0.69	**0.61**	0.59	0.68	0.63	0.63	0.38	0.47	0.59	0.58	**0.57**
	R→D	0.43	0.78	0.55	0.55	0.49	0.52	0.70	0.20	0.31	0.56	0.49	0.46
Qwen2.5	D	0.52	0.67	0.59	0.56	**0.82**	**0.66**	**0.96**	0.24	0.38	**0.68**	0.58	0.54
	R→D	0.40	0.84	0.54	0.55	0.39	0.45	0.94	0.18	0.30	0.63	0.47	0.43
DeepSeek	D	0.52	0.43	0.47	0.48	0.72	0.57	0.41	0.27	0.33	0.47	0.47	0.46
	R→D	0.40	0.60	0.48	0.45	0.50	0.47	0.43	0.16	0.23	0.42	0.42	0.39

Table 8: Model performance on the claim verification task of MuSciClaims for Physics claims when prompted to simply generate the decision (D), and when asked to reason and then generating the decision (R→D).

the scientific field.

2. Visual Understanding (23%) - Models are unable to make the correct inference from the information that they perceive from the figure.
3. Visual Perception (23%) - Models either miss or pick up data presented in the figure, which can lead to the correct judgment.
4. Cross Modal Aggregation (17%) - Models incorrectly weight information from one modality over another (such as focus on the caption more than the figure), coming to the wrong conclusion.
5. Others (10%) - This category contains infrequent error types bucketed together. For instance, models simply misunderstand the caption (textual understanding), or fail to aggregate information from multiple panels (multi-panel aggregation).

G Prompts Used

We present the exact prompts used for different experiments with Sonnet in Figure 8, Figure 9 and Figure 10 and InternVL3 in Figure 11, Figure 12 and Figure 13.

		SUPPORT			NEUTRAL			CONTRADICT			OVERALL		
		P	R	F	P	R	F	P	R	F	P	R	F
4o-mini	D	0.46	0.86	0.60	0.70	0.69	0.69	0.79	0.11	0.19	0.65	0.55	0.49
	R→D	0.46	0.74	0.57	0.63	0.67	0.65	0.63	0.21	0.32	0.57	0.54	0.51
4o	D	0.46	0.91	0.61	0.91	0.59	0.72	0.70	0.25	0.37	0.69	0.59	0.57
	R→D	0.49	0.82	0.61	0.76	0.71	0.73	0.66	0.26	0.38	0.64	0.60	0.57
Sonnet	D	0.49	0.85	0.63	0.89	0.62	0.73	0.66	0.38	0.48	0.68	0.62	0.61
	R→D	0.53	0.87	0.66	0.90	0.63	0.74	0.74	0.48	0.58	0.72	0.66	0.66
o3	R→D	**0.75**	0.66	0.70	0.67	0.86	0.76	0.73	0.61	**0.67**	0.72	0.71	0.71
o4-mini	R→D	0.67	0.78	**0.72**	0.72	0.82	0.76	0.81	0.56	0.66	**0.73**	**0.72**	0.71
Phi-4	D	0.43	0.71	0.54	**0.97**	0.29	0.45	0.41	0.44	0.42	0.60	0.48	0.47
	R→D	0.35	0.86	0.50	0.56	0.05	0.09	0.65	0.27	0.38	0.52	0.39	0.32
Llava-Next	D	0.39	0.90	0.55	0.68	0.47	0.55	0.00	0.00	0.00	0.36	0.46	0.37
	R→D	0.38	0.80	0.51	0.60	0.45	0.51	0.69	0.11	0.18	0.55	0.45	0.40
Llama-3.2	D	0.40	0.86	0.55	0.81	0.38	0.52	0.46	0.18	0.26	0.56	0.48	0.44
	R→D	0.38	0.93	0.54	0.67	0.12	0.20	0.53	0.18	0.27	0.52	0.41	0.34
Molmo	D	0.42	**0.92**	0.58	0.78	0.27	0.40	0.44	0.20	0.28	0.55	0.47	0.42
	R→D	0.38	0.80	0.52	0.58	0.28	0.38	0.36	0.16	0.22	0.44	0.41	0.37
InternVL3	D	0.66	0.76	0.70	0.88	0.74	0.80	0.65	**0.66**	0.66	**0.73**	**0.72**	**0.72**
	R→D	0.44	0.89	0.59	0.96	0.53	0.69	0.66	0.28	0.39	0.69	0.57	0.56
Qwen2.5	D	0.56	0.74	0.64	0.68	**0.90**	**0.78**	**0.86**	0.31	0.46	0.70	0.65	0.62
	R→D	0.41	0.91	0.57	0.88	0.43	0.58	0.81	0.24	0.37	0.70	0.53	0.51
DeepSeek	D	0.62	0.36	0.45	0.50	0.76	0.60	0.43	0.39	0.41	0.52	0.50	0.49
	R→D	0.41	0.69	0.52	0.53	0.54	0.54	0.44	0.14	0.21	0.46	0.46	0.42

Table 9: Model performance on the claim verification task of MUSCICLAIMS for CHEMISTRY claims when prompted to simply generate the decision (D), and when asked to reason and then generating the decision (R→D).

		SUPPORT			NEUTRAL			CONTRADICT			OVERALL		
		P	R	F	P	R	F	P	R	F	P	R	F
4o-mini	D	0.40	0.90	0.56	0.66	0.42	0.51	0.74	0.10	0.18	0.60	0.47	0.42
	R→D	0.42	0.87	0.57	0.68	0.41	0.51	0.62	0.21	0.31	0.57	0.50	0.46
4o	D	0.43	0.96	0.59	0.93	0.44	0.60	0.84	0.23	0.36	0.73	0.54	0.52
	R→D	0.48	0.92	0.63	0.76	0.58	0.66	0.83	0.26	0.39	0.69	0.59	0.56
Sonnet	D	0.53	0.92	0.67	0.91	0.63	0.74	0.82	0.49	0.61	0.76	0.68	0.68
	R→D	0.55	0.92	0.69	0.94	0.64	0.76	0.84	0.54	0.65	0.78	0.70	0.70
o3	R→D	**0.67**	0.84	**0.75**	0.80	**0.77**	0.78	0.89	0.69	**0.78**	0.79	**0.77**	**0.77**
o4-mini	R→D	0.63	0.92	**0.75**	0.87	0.75	**0.80**	0.92	0.64	0.76	**0.81**	**0.77**	**0.77**
Phi-4	D	0.46	0.69	0.55	**0.98**	0.14	0.24	0.44	0.60	0.51	0.63	0.47	0.43
	R→D	0.36	0.86	0.51	0.85	0.13	0.23	0.57	0.27	0.37	0.60	0.42	0.37
Llava-Next	D	0.37	**0.98**	0.53	0.85	0.27	0.41	**1.00**	0.01	0.01	0.74	0.42	0.32
	R→D	0.39	0.82	0.53	0.58	0.42	0.49	0.55	0.08	0.14	0.50	0.44	0.38
Llama-3.2	D	0.42	0.88	0.57	0.70	0.44	0.54	0.67	0.18	0.29	0.60	0.50	0.47
	R→D	0.38	0.95	0.54	0.79	0.16	0.27	0.63	0.17	0.27	0.60	0.43	0.36
Molmo	D	0.42	0.92	0.57	0.85	0.32	0.47	0.60	0.24	0.34	0.62	0.50	0.46
	R→D	0.39	0.75	0.51	0.57	0.30	0.39	0.44	0.25	0.32	0.47	0.43	0.41
InternVL3	D	0.63	0.84	0.72	0.93	0.66	0.77	0.74	**0.71**	0.72	0.77	0.74	**0.74**
	R→D	0.46	0.96	0.62	0.93	0.47	0.62	0.88	0.35	0.50	0.75	0.59	0.58
Qwen2.5	D	0.55	0.85	0.67	0.77	0.75	0.76	0.83	0.37	0.51	0.72	0.66	0.65
	R→D	0.43	0.93	0.59	0.86	0.42	0.56	0.84	0.28	0.43	0.71	0.54	0.52
DeepSeek	D	0.55	0.45	0.49	0.52	0.63	0.57	0.46	0.44	0.45	0.51	0.51	0.50
	R→D	0.42	0.66	0.51	0.53	0.50	0.51	0.54	0.26	0.35	0.50	0.47	0.46

Table 10: Model performance on the claim verification task of MUSCICLAIMS for BIOLOGY claims when prompted to simply generate the decision (D), and when asked to reason and then generating the decision (R→D).

		SUPPORT			NONSUPPORT			OVERALL		
		P	R	F	P	R	F	P	R	F
4o-mini	D	0.39	0.93	0.55	0.89	0.29	0.43	0.72	0.5	0.47
	R→D	0.41	0.92	0.56	0.89	0.33	0.49	0.73	0.53	0.51
	I→R→D	0.37	0.96	0.53	0.91	0.18	0.30	0.73	0.44	0.37
4o	D	0.41	0.96	0.57	0.94	0.30	0.46	0.77	0.52	0.50
	R→D	0.43	0.95	0.59	0.94	0.38	0.54	0.77	0.57	0.56
	I→R→D	0.39	0.98	0.56	0.97	0.25	0.39	0.78	0.49	0.45
Sonnet	D	0.52	0.93	0.67	0.94	0.57	0.71	0.80	0.69	0.70
	R→D	0.53	0.95	0.68	0.96	0.58	0.72	0.82	0.70	0.71
	I→R→D	0.51	0.96	0.66	0.96	0.53	0.69	0.81	0.68	0.68

Table 11: Model performance on BIOLOGY claims when posing MUSCICLAIMS as a two class problem.

		Single-Panel			Multi-Panel		
		P	R	F	P	R	F
4o-mini	D	0.59	0.47	0.41	0.65	0.51	0.47
	R→D	0.58	0.49	0.46	0.56	0.51	0.48
	I→R→D	0.6	0.45	0.4	0.57	0.46	0.41
4o	D	0.73	0.54	0.52	0.77	0.55	0.53
	R→D	0.69	0.58	0.56	0.72	0.61	0.59
	I→R→D	0.73	0.51	0.47	0.76	0.53	0.5
Sonnet	D	0.75	0.68	0.68	0.78	0.68	0.69
	R→D	0.78	0.7	0.7	0.79	0.72	0.72
	I→R→D	0.79	0.7	0.7	0.76	0.67	0.68

Table 12: Model performance on BIOLOGY claims when broken down by complexity of visual aggregation for claims

Dataset	Scientific?	Multimodal?	Claim Verification?	Heterogeneous Figures?	Real?	Complex?
ArXivQa	✓	✓	×	✓	✓	×
MMC	✓	✓	×	✓	✓	×
PlotQA	×	✓	×	×	×	×
SPIQA	✓	✓	×	✓	✓	×
FigureQA	×	✓	×	×	×	×
DVQA	×	✓	×	×	×	×
ChartQA	×	✓	×	×	✓	×
ChartBench	×	✓	×	✓	✓	×
ChartX	×	✓	×	✓	×	×
MultiChartQA	×	✓	×	✓	✓	✓
ChartCheck	×	✓	✓	×	✓	×
SciFact	✓	×	✓	×	×	×
SciFact-Open	✓	×	✓	×	×	×
PubHealthTab	×	×	✓	×	×	×
MuSciClaims	✓	✓	✓	✓	✓	✓

Table 13: Comparison of MUSCICLAIMS against related work benchmarks across different desired characteristics for a multimodal claim verification task.

You are an AI model tasked with verifying claims related to visual evidence using zero-shot learning. Your job is to analyze a given image(s) and its provided caption(s) to decide whether it SUPPORT or CONTRADICT or NEUTRAL the provided claim.

CLAIM: CLAIM
IMAGE CAPTION(S): IMAGE_CAPTIONS

Guidelines:
1. Evaluate the claim's plausibility based on visual elements within the image(s).
2. Consider the relevance, meaning, and implications of both the depicted content and the caption(s).
3. Analyze the broader context and scope of the image(s) and caption(s) in relation to the claim.

After completing your analysis, output exactly one JSON object with exactly one key: "decision".
- For "decision", output exactly one word — either "SUPPORT" or "CONTRADICT" or "NEUTRAL" (uppercase, no extra text).
Do NOT add markdown formatting, code fences, or any additional text. The output must start with an opening curly brace { and end with a closing curly brace }.

Example output format:
{"decision": "SUPPORT"}

Now, please evaluate the image(s) and caption(s) with respect to the claim provided above.

Figure 8: Prompt for `Sonnet` for the D experiment

You are an AI model tasked with verifying claims related to visual evidence using zero-shot learning. Your job is to analyze a given image(s) and its provided caption(s) to decide whether it SUPPORT or CONTRADICT or NEUTRAL the provided claim.
CLAIM: {CLAIM}
IMAGE CAPTION(S): {IMAGE_CAPTIONS}
Guidelines:
1. Evaluate the claim's plausibility based on visual elements within the image(s). 2. Consider the relevance, meaning, and implications of both the depicted content and the caption(s). 3. Analyze the broader context and scope of the image(s) and caption(s) in relation to the claim. 4. Think step by step to reach your conclusion, but only provide a concise reasoning statement in the output.
After completing your analysis, output exactly one JSON object with exactly two keys in this order: "reasoning" and "decision".
- For "reasoning", provide a brief (one- or two-sentence) explanation of your analysis.
- For "decision", output exactly one word — either "SUPPORT" or "CONTRADICT" or "NEUTRAL" (uppercase, no extra text).
Do NOT add markdown formatting, code fences, or any additional text. The output must start with an opening curly brace { and end with a closing curly brace }.
Example output format:
{"reasoning": "The caption confirms the rising trend visible in the image, supporting the claim.", "decision": "SUPPORT"}
Now, please evaluate the image(s) and caption(s) with respect to the claim provided above.

Figure 9: Prompt for `Sonnet` for the R→D experiment

You are an AI model tasked with verifying claims related to visual evidence using zero-shot learning. Your job is to analyze a given image(s) and its provided caption(s) to decide whether it SUPPORT or CONTRADICT or NEUTRAL the provided claim.

CLAIM: CLAIM
IMAGE CAPTION(S): IMAGE_CAPTIONS

Guidelines:
1. Evaluate the claim's plausibility based on visual elements within the image(s).
2. Consider the relevance, meaning, and implications of both the depicted content and the caption(s).
3. Analyze the broader context and scope of the image(s) and caption(s) in relation to the claim.
4. Identify which specific panels (e.g., Panel A, Panel B, Panel C, etc.) are necessary to evaluate the claim.
5. Think step by step to reach your conclusion and provide it in a concise manner in the output.

After completing your analysis, output exactly one JSON object with exactly three keys in this order: "figure_panels", "reasoning", and "decision".
- For "figure_panels", list ONLY the names or labels of the panels needed to evaluate the claim (e.g., ["Panel A", "Panel C"]) with no further description. If no panels are needed, return [].
- For "reasoning", provide a brief (one- or two-sentence) explanation of your analysis.
- For "decision", output exactly one word — either "SUPPORT" or "CONTRADICT" or "NEUTRAL" (uppercase, no extra text).
Do NOT add markdown formatting, code fences, or any additional text. The output must start with an opening curly brace { and end with a closing curly brace }.

Example output format:
{"figure_panels": ["Panel A", "Panel C"], "reasoning": "The trend in Panel A aligns with the claim, while Panel C corroborates the effect.", "decision": "SUPPORT"}

Now, please evaluate the image(s) and caption(s) with respect to the claim provided above.

Figure 10: Prompt for `Sonnet` for the I→R→D experiment

This is an image from a scientific paper. The following is the caption of the image.

IMAGE CAPTION(S): IMAGE_CAPTIONS

Using this image, analyze whether the following claim is supported, contradicted or neutral according to the image and caption.

CLAIM: CLAIM

Reply with one of the following keywords: SUPPORT, CONTRADICT, NEUTRAL. Do not generate any other text or explanation.

Return your answer in following format:
DECISION: <your decision>

Figure 11: Prompt for `InternVL3` for the D experiment

This is an image from a scientific paper. The following is the caption of this image.

IMAGE CAPTION(S): IMAGE_CAPTIONS

Using this image, analyze whether the following claim is supported, contradicted or neutral according to the image and caption.

CLAIM: CLAIM

Think step by step to reach your conclusion and then reply with only one of the following keywords: SUPPORT, CONTRADICT, NEUTRAL. Your reasoning should be brief and concise, no more than 100 words.

Return your answer in following format:
REASONING: <your reasoning>
DECISION: <your decision>

Figure 12: Prompt for `InternVL3` for the R→D experiment

This is an image, with multiple panels, from a scientific paper. The following is the caption of this image.

IMAGE CAPTION(S): IMAGE_CAPTIONS

Using this image, analyze whether the following claim is supported, contradicted or neutral according to the image and caption.

CLAIM: CLAIM

First identify the relevant panels (Figure A, Figure B etc.) in the image that are needed to analyze the claim. Then think step by step to reach your conclusion and reply with only one of the following keywords: SUPPORT, CONTRADICT, NEUTRAL. Your reasoning should be brief and concise, no more than 100 words.

Return your answer in following format:
FIGURE PANELS: <the figure panels to use for deduction>
REASONING: <your reasoning>
DECISION: <your decision>

Figure 13: Prompt for `InternVL3` for the I→R→D experiment