

STUDENT GENG'S WORKBUDDY: EVIDENCE-GROUNDED VERIFICATION OF SCIENTIFIC CONCERNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post publication peer review supports scientific self-correction, but systematic, reproducible verification of public scientific concerns still relies heavily on manual effort. PubPeer is an open platform where researchers question published papers, share supporting material, and discuss the reliability of the work. It has consequently become a large collection of research concerns. Recently, Hongwei Geng, also known as Student Geng, used public PubPeer comments and related evidence to examine data anomalies in papers from leading journals, such as *Nature*. The resulting cases led to investigations and sanctions by organizations such as the National Natural Science Foundation of China. Such cases show that PubPeer comments provide scientifically and practically valuable leads for research scrutiny, but remain unverified and require substantial manual validation. To address this problem, we propose Student Geng's WorkBuddy-27B, an evidence verification assistant for PubPeer comments. For each comment and corresponding paper, WorkBuddy first reformulates the scientific concern as a neutral statement, removing subjective or emotional language while preserving verifiable factual claims. It then maps the concern to traceable evidence in paper text, figures, methods, supplementary materials, and source data. WorkBuddy assesses how strongly this evidence supports the concern, together with evidence quality, potential impact, and evidence location. For training and evaluation, we construct PubPeerBench-1000 with 1,000 cases annotated by experts. We train Student Geng's WorkBuddy through supervised fine-tuning and compare it on an independent test set with leading large language models, including GPT-6 Astra. Against untuned Qwen3.6-27B, our model's verdict exact agreement rises from 38.50% to 48.00%, Macro-F1 from 28.13% to 37.00%, EQ MAE falls from 1.50 to 0.89, and impact exact agreement rises from 23.50% to 39.50%, showing the effectiveness of task specific fine tuning. *PubPeerBench-1000 and code will be publicly released.*

1 INTRODUCTION

Questions about data anomalies, duplicated images, and overextended conclusions have remained central to discussions of research integrity (Bik et al., 2016; Fanelli, 2009; Head et al., 2015; Ioannidis, 2005). Once a paper is published, checking such problems is part of the work of maintaining a reliable record. In 2026, *Science* reported widespread attention to Hongwei Geng, an independent video blogger in China who publicly identified suspicious papers (Normile, 2026). *Nature* later reported that related allegations of data anomalies led to institutional investigations and disciplinary action against four senior Chinese scholars (You, 2026). Chinese research management agencies have also increased their attention to serious research misconduct and institutional accountability (You, 2026). Many of Geng's leads came from post publication peer review platforms such as PubPeer (You, 2026; Townsend, 2013; Ortega, 2022). A PubPeer comment raises a question or points to material for investigation (Townsend, 2013; Ortega, 2022). It is not normally verified beforehand and does not amount to a formal misconduct finding. Our task is therefore narrower: determine whether the evidence in a target paper supports a particular concern raised in a comment. Figure 1 summarizes this motivation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

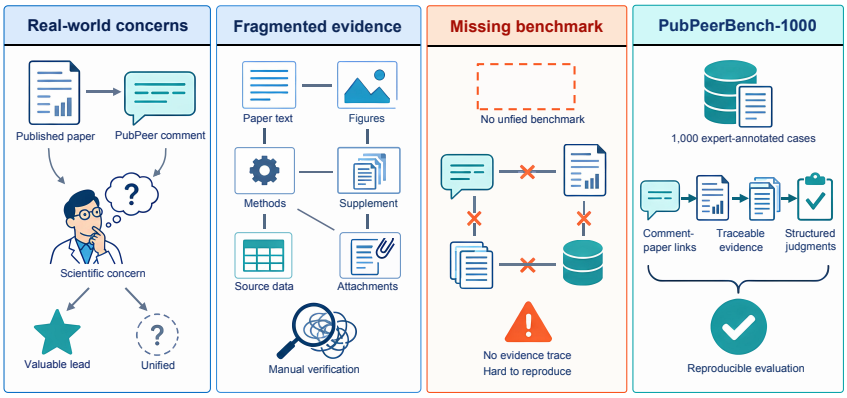

Figure 1: From public scientific concerns to evidence based verification. PubPeer comments can identify valuable leads, but the relevant evidence is distributed across paper text, figures, methods, supplementary materials, data, and attachments. PubPeerBench-1000 links each comment to its target paper, traceable evidence, and structured expert judgments to support reproducible evaluation.

Computational work on possible problems in scientific papers includes text similarity detection, image forensics, statistical and data forensics, author and citation network analysis, and post publication peer review. Such methods can flag unusual text, images, data, or research networks and can produce leads for further checking (Bik et al., 2016; Nuijten et al., 2015; Fang et al., 2012; Steen, 2010). They usually do not answer the more specific question of whether the evidence in a given paper supports a particular concern. That requires linking the concern to the relevant paper content and then judging the strength of the support.

We construct PubPeerBench-1000 after obtaining permission. The benchmark contains 1,000 comment verification cases covering 357 DOIs, with each case pairing one PubPeer comment with the paper it refers to. From the complete paper JSON, we extract the main text, figures, tables, captions, and methods relevant to the comment, retain image and page location metadata, and package the result as an evidence JSON bundle related to the claim. Two experts provide independent annotations for every case, and a third expert resolves disagreements. The final verdict labels are `supported`, `partially_supported`, `unsupported`, and `insufficient_evidence`. We also record evidence quality, impact, and a rationale. Missing necessary evidence is kept separate from evidence that directly refutes a concern.

We use PubPeerBench-1000 to train Student Geng's WorkBuddy-27B. Starting from Qwen3.6-27B (Qwen Team, 2026), we apply LoRA supervised fine tuning with a concern and its matched paper materials as input. The model predicts evidence support, evidence quality, and impact. We evaluate it against the base model and several leading general purpose models.

On the same 200 test cases grouped by paper association and with identical input materials, Student Geng's WorkBuddy-27B raises verdict exact agreement from 38.50% to 48.00%, Macro-F1 from 28.13% to 37.00%, and impact exact agreement from 23.50% to 39.50%. EQ MAE falls from 1.50 to 0.89. The trained model also exceeds GPT-6 Astra (OpenAI, 2026) on the reported metrics. In this setting, task specific fine tuning improves verification performance.

Our contributions are threefold:

1. We construct PubPeerBench-1000 with 1,000 cases and traceable paper text and image evidence, giving evidence verification of PubPeer concerns a common dataset and evaluation basis.

2. We formulate comment verification as a multidimensional evidence judgment task that reports support, evidence quality, and impact separately.

3. We train and evaluate Student Geng's WorkBuddy-27B for PubPeer verification and examine its behavior alongside a base model and leading general purpose models.

2 Related Work

Computational studies of possible problems in scientific papers address text, images, data, and networks of research relationships. Text similarity and plagiarism systems use lexical, syntactic, or semantic matching to find copying, paraphrase, and text reuse, so their focus is on source overlap (Beltagy et al., 2019; Reimers & Gurevych, 2019; Khattab & Zaharia, 2020). Image forensics methods detect duplication, rotation, splicing, and local copying, but they usually do not connect a detected anomaly to a specific scientific concern (Bik et al., 2016; Bucci, 2018; Sabir et al., 2021; Xiang & Acuna, 2020; Cicconet et al., 2018; Hao et al., 2021; Bravo-Solorio & Nandi, 2011; Christlein et al., 2012; Zhao & Zhao, 2013). Statistical and data forensics methods examine distributions, numerical patterns, statistical tests, and reporting consistency. Many require raw data and domain expertise (Nuijten et al., 2015; Head et al., 2015; Fanelli, 2009; Ioannidis, 2005). Author collaboration networks, citation links, and publication patterns can expose features associated with paper mills or organized research activity and are commonly used for large scale screening and ranking. Platforms such as PubPeer collect concerns about published papers and can provide early warnings, while leaving formal evidence verification to later investigation (Byrne et al., 2024; Fang et al., 2012; Steen, 2010; Townsend, 2013; Ortega, 2022; Caron et al., 2024).

These methods mainly flag anomalies and produce leads for follow up. They do not directly measure how strongly a particular concern is supported by its target paper. Our method links each concern to relevant text, figures, and methodological details, then checks it against multimodal evidence so that the resulting judgment is specific and traceable (Wadden et al., 2020; 2022; Sarrouti et al., 2021; Jiang et al., 2020; Aly et al., 2021; Atanasova et al., 2020; Kotonya & Toni, 2020).

3 Method

We build PubPeerBench-1000 to test whether a model can check scientific concerns in PubPeer comments against paper evidence. The workflow covers comment collection, neutralized concern reconstruction, paper material parsing, evidence bundle construction, expert annotation, splitting by paper association groups, and model training, as shown in Figure 2. Versioned JSON stores the cases, evidence, and labels, making the source materials, processing steps, and judgments traceable.

3.1 Task Definition and Input and Output Representation

We treat PubPeer comment verification as structured prediction grounded in paper evidence. The basic unit is a comment paired with the paper it names. A comment can contain several independently checkable subclaims, but the final label is assigned to the case and summarizes support for its key claims. The task does not decide whether a paper is fraudulent, infer author intent, assign legal responsibility, or make a formal misconduct finding.

Each case stores a unique `case_id`, paper title and DOI, the original comment, a neutralized concern `claim`, multi label issue types `issue_types`, evidence scope `evidence_scope`, the complete paper JSON, comment attachments, and an evidence JSON bundle related to the claim. Model inputs contain the comment or neutralized concern, matched evidence text, and the corresponding images. We record text, claim only, original comment, and mismatched evidence conditions separately. The complete paper JSON supports archival traceability. It is not the same as the model input.

The model returns four structured fields. `verdict` records how far the available evidence supports the concern. `evidence_quality` describes the relevance, reproducibility, and completeness of that evidence. `impact` records the possible effect on the paper's results or interpretation if the concern is true. `rationale` briefly describes the observed materials, their connection to the concern, and any remaining limitations.

3.2 Comment Collection and Case Construction

After obtaining permission, we collected visible comments from the PubPeer Selected feed and discussion pages. A candidate needed a complete comment identity and number, an explicit DOI, a nonempty body, visible and enabled status, and an ordinary commenter role rather than an au-

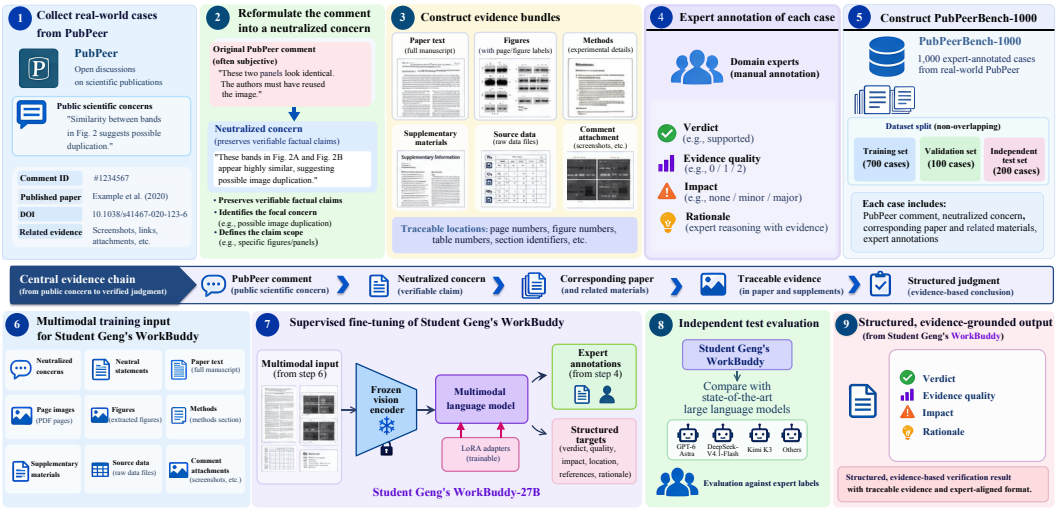

Figure 2: Overview of the PubPeerBench-1000 construction and Student Geng's WorkBuddy training pipeline. The workflow collects real world PubPeer cases, reconstructs neutralized concerns, assembles traceable evidence bundles, obtains expert annotations, and builds the benchmark. The resulting multimodal inputs are used for supervised fine tuning and independent test evaluation, producing structured outputs grounded in paper evidence.

thor reply. We randomly sampled 1,000 comments covering 357 DOIs, as shown in Figure 3. Each record links back to the original data through `case_id`, publication URL, comment number, `comment_id`, and DOI. This identity chain supports deduplication, paper matching, and reproduction. In compact form, the chain is `case_id` → `publication_url` → `comment_number` → `comment_id` → DOI.

3.3 Neutralized Concern Reconstruction

An original comment may mix figure comparisons, numerical checks, experimental methods, statistical analyses, alternative explanations, and author replies. We reconstruct `claim` from the complete comment as a neutralized concern that can be matched to paper evidence. The reconstruction keeps the main problem, comparison target, key numbers, figure and table identifiers, statistical procedures, methodological steps, and independent substantive subclaims. It does not reduce the comment to a single overall judgment.

The reconstruction keeps epistemic qualifiers such as "may", "appears" and "requires checking". Phrases such as "the comment points out", "the comment questions" and "the comment recommends checking" distinguish the commenter's view from an established fact in the paper. Terms such as "fabrication", "deliberately misleading" or "research misconduct" remain attributed to the comment. They are not turned into dataset facts or expert conclusions. We do not infer the contents of unseen attachments or add problems that the comment does not raise. Author replies and cited material remain separate materials.

3.4 JSON Evidence Bundle Construction

Each case is stored as a structured evidence JSON bundle related to the claim, with paper materials, comment materials, and attachments. We match the paper by DOI and parse it into page- and structure aware JSON containing the main text, figures, tables, captions, methods, supplementary materials, source URLs, parsing status, and content hashes.

The evidence JSON bundle related to the claim is retrieved by location first and then by text anchors (Karpukhin et al., 2020; Lewis et al., 2021; Guu et al., 2020; Izacard & Grave, 2021). The system uses page numbers, figure and table identifiers, and section information in the comment to extract the corresponding objects. If no explicit location is given, it searches page-level text for technical

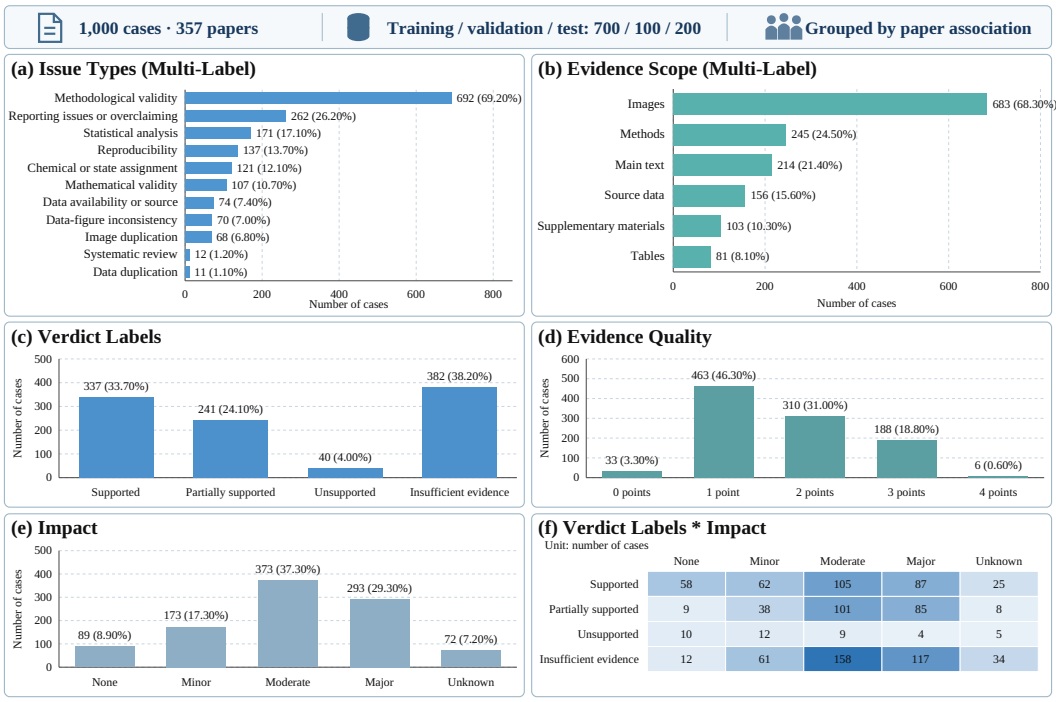

Figure 3: Distribution of PubPeerBench-1,000 cases across issue types, evidence scopes, verdict labels, evidence quality scores, and impact levels. Panels a and b show the distributions of issue types and evidence scopes, with multiple labels allowed for each case. Panels c to e report the distributions of verdict labels, evidence quality scores, and impact levels. Panel f shows the joint distribution of verdict labels and impact. The benchmark contains 1,000 cases from 357 papers, with 700 training cases, 100 validation cases, and 200 test cases.

terms, key numbers, and key phrases. For a figure related concern, the bundle keeps the relevant figure, caption, or table note with its context. For a method or interpretation concern, it keeps the relevant text and method descriptions. Each record stores text blocks, figure or table objects, the matching rationale, retrieval status, and a short summary. Image objects point to the actual image files, and multimodal conditions load those files instead of a path, URL, or textual description.

3.5 Expert Annotation and Quality Control

Verdict has four classes. `supported` means that the available materials support the key claims. `partially_supported` means that they support some key subclaims but not the concern as a whole. `unsupported` means that the materials provide counter evidence or that an explanation already present in the paper resolves the concern. `insufficient_evidence` means that key materials are missing or that the evidence cannot support or refute the concern. A lack of supporting evidence is not automatically `unsupported`.

Evidence quality uses an ordered score from 0 to 4 based on relevance, locatability, reproducibility, and source consistency. A score of 0 means that no evidence can be evaluated. 1 means that the material is indirect or hard to locate. 2 means that an observable phenomenon is present but the materials are incomplete. 3 means that the evidence can be checked independently from the available materials. And 4 means that several sources provide consistent support. Impact uses `none`, `minor`, `moderate`, `major`, and `unknown`. These levels correspond to no substantive problem, a local reporting issue, a secondary result or local interpretation, a main result or core method, and an indeterminate effect. The rationale states the observed evidence, its relation to the concern, and any alternative explanations or material limitations in one to three sentences.

Two experts annotate each case independently. Before adjudication, agreement between the two experts is measured on all 1,000 cases. The resulting Cohen's kappa values are 0.89 for verdict and 0.92 for impact. Evidence quality is recorded on an ordinal scale, so weighted Cohen's kappa is used, with a value of 0.81. A third expert resolves disagreements and produces the final expert labels, while the independent annotations and adjudication rationale are retained. The adjudication process covers verdict disagreements, differences in evidence quality or impact levels, and substantive conflicts about key evidence. Quality control checks the case count, identity chain, matching between DOIs and paper JSON records, record hashes, valid field values, retrieval states, independent annotations, and adjudication records. Formal evaluation uses the final expert labels.

3.6 Data Splitting and Leakage Control

We form paper association groups from DOI, paper version relations, cross paper links, and near duplicate signals. We then split the data into 700 training, 100 validation, and 200 test cases. No association group crosses subsets, which reduces the chance that the same paper text, images, or evidence materials appear in both training and test data.

3.7 Student Geng's WorkBuddy

We use Qwen3.6-27B as the base model and apply LoRA-SFT to obtain Student Geng's WorkBuddy-27B (Hu et al., 2021; Dettmers et al., 2023). Training inputs contain the neutralized concern and matching evidence JSON bundle related to the claim. Multimodal conditions also load the matched images (Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023; Liu et al., 2023; Bai et al., 2023; Peng et al., 2023; Chen et al., 2022; Jaegle et al., 2021; Dosovitskiy et al., 2021; Radford et al., 2021). The targets are the final expert labels for `verdict`, `evidence_quality`, `impact`, and `rationale`. Training uses BF16, freezes the base model, vision encoder, and multimodal projection, and updates LoRA adapters in the attention and MLP layers of the language module. We evaluate verdict exact agreement, verdict Macro-F1, EQ MAE, and impact exact agreement. On the same 200 test cases, Student raises verdict exact agreement from 38.50% to 48.00%, Macro-F1 from 28.13% to 37.00%, and impact exact agreement from 23.50% to 39.50%. EQ MAE falls from 1.50 to 0.89. All four reported metrics improve.

4 Experiments

4.1 Experimental Setup

All experiments use the fixed 200-case test set, the same output fields, and the final expert labels. The data are split by paper association groups into 700 training cases, 100 validation cases, and 200 test cases, with no group crossing subsets. Unless noted otherwise, models receive a neutralized concern together with matched paper text and image evidence. The output contains `verdict`, `evidence_quality`, `impact`, and `rationale`. The first three fields are evaluated quantitatively, while the rationale records the basis for the judgment.

Verdict has four classes: `supported`, `partially_supported`, `unsupported`, and `insufficient_evidence`. We evaluate it with exact agreement and Macro-F1. Exact agreement is the proportion of predictions that exactly match the expert label, and Macro-F1 is the unweighted mean of the F1 scores over the four classes. Impact is evaluated by exact agreement, with `unknown` treated as an independent state rather than as a severity level. Evidence quality uses the ordered scores from 0 to 4 and is evaluated with EQ MAE, the mean absolute difference between the model score and the expert score. The difference is computed on the numeric encodings. This is an evaluation convention and does not imply that adjacent quality levels represent equal real world differences. Lower EQ MAE is better, while higher values are better for the other metrics.

4.2 Task Fine tuning and General model Comparison

We compare untuned Qwen3.6-27B, Student Geng's WorkBuddy-27B, GPT-6 Astra, Qwen3.8-Max, DeepSeek-V4.1-Flash, and Kimi K3 (OpenAI, 2026; Alibaba Cloud, 2026; DeepSeek-AI, 2026;

Table 1: Model performance on PubPeerBench-1000. Up arrows indicate that higher values are better. Down arrows indicate that lower values are better.

Model	Verdict exact agreement ↑	Macro-F1 ↑	EQ MAE ↓	Impact exact agreement ↑
GPT-6 Astra	46.50%	34.38%	0.98	33.50%
Qwen3.8-Max	42.50%	33.13%	1.04	30.00%
DeepSeek-V4.1-Flash	44.00%	31.50%	1.17	28.50%
Kimi K3	41.00%	30.62%	1.11	27.00%
Qwen3.6-27B	38.50%	28.13%	1.50	23.50%
Student Geng's WorkBuddy	**48.00%**	**37.00%**	**0.89**	**39.50%**

Moonshot AI, 2026; Qwen Team, 2026). Every model receives the same neutralized concern and matched paper evidence.

Table 1 shows the effect of task specific fine tuning. Compared with the base model, Student Geng's WorkBuddy-27B improves on all four evaluation metrics. Verdict exact agreement increases from 38.50% to 48.00%. Macro-F1 increases from 28.13% to 37.00%, giving gains of 9.50 and 8.87 percentage points, respectively. Impact exact agreement increases from 23.50% to 39.50%. EQ MAE decreases from 1.50 to 0.89, a relative reduction of 40.66%. These results show that, on the current test set, task specific fine tuning improves agreement with the final expert labels for verdict and impact while reducing the average level error in evidence quality scores.

Among the three general purpose models, GPT-6 Astra performs better on all four metrics than Kimi K3, DeepSeek-V4.1-Flash, and Qwen3.8-Max. We therefore use it as the main cross-model comparison. Compared with GPT-6 Astra, Student Geng's WorkBuddy-27B improves verdict exact agreement, Macro-F1, and impact exact agreement by 1.50, 2.62, and 6.00 percentage points, respectively. Its EQ MAE is lower by 0.09. In this evaluation, Student Geng's WorkBuddy-27B outperforms all compared general purpose models on all four metrics. These results support using Student Geng's WorkBuddy-27B to verify scientific concerns raised on PubPeer.

4.3 Number of Issues and Verification Performance

This experiment examines the relationship between the number of independently verifiable issues in a comment and model verification performance, with a focus on whether verdict exact agreement decreases monotonically as the issue count increases. We divide the fixed 200 test cases from PubPeerBench-1000 into four groups according to issue count: 0, 1, 2, and at least 3 independently verifiable issues. The groups contain 23, 88, 41, and 48 cases, respectively. "0 issues" means that the current issue decomposition procedure did not extract a clear, independently verifiable scientific subproblem. Such a comment may mainly provide background information or general discussion. This label does not mean that the comment raises no concern or that the target paper has no problems. We report the verdict exact agreement of the Qwen3.6-27B base model, Student Geng's WorkBuddy-27B, Kimi K3, and DeepSeek-V4.1-Flash separately for each issue count group.

Table 2: Verdict exact agreement by the number of independently verifiable issues.

Model	0 issues	1 issue	2 issues	$\geq$3 issues
Kimi K3	26.08%	43.18%	51.21%	41.67%
DeepSeek-V4.1-Flash	34.78%	46.59%	53.66%	45.83%
Qwen3.6-27B	21.73%	40.91%	43.90%	37.50%
Student Geng's WorkBuddy	**30.43%**	**47.73%**	**58.54%**	**47.92%**

Table 2 reports the stratified results. The verdict exact agreement of all four models first increases and then decreases rather than declining monotonically as the issue count grows. For Student Geng's WorkBuddy-27B, verdict exact agreement rises from 30.43% in the 0 issues group to 47.73% in the

1 issue group, reaches 58.54% in the 2 issues group, and then falls to 47.92% in the at least 3 issues group. Even after excluding the 0 issues cases, whose meaning differs from that of the other groups, the remaining three groups still show a rise followed by a fall. This observation is therefore not caused solely by including the 0 issues group.

The same pattern appears in the base model and the two general purpose models. Qwen3.6-27B, Kimi K3, and DeepSeek-V4.1-Flash all achieve their highest verdict exact agreement in the 2 issues group, with values of 43.90%, 51.21%, and 53.66%, respectively. Their verdict exact agreement then decreases in the at least 3 issues group. The nonmonotonic pattern is therefore not unique to Student Geng's WorkBuddy-27B. It can be observed in all four evaluated models.

In the current test set and under this grouping scheme, verdict exact agreement does not decline monotonically as the number of independently verifiable issues increases. This result does not support equating a larger issue count directly with worse model verification performance. It also suggests that issue count alone should not be used to judge case verification difficulty. More generally, issue count and model verification performance do not have a simple linear relationship, so issue count alone is insufficient to predict verification outcomes.

4.4 Effect of Comment Wording

We run a paired comparison on the same 200 test cases. The `original` condition uses the complete PubPeer comment, including its wording, tone, and line of argument, whereas `neutral-multimodal` uses the neutralized concern extracted from that comment. Model weights, instructions, paper text, image evidence, and decoding configuration stay fixed, so only the comment input changes. Both outputs are compared with the same final expert labels, using verdict exact agreement, Macro-F1, EQ MAE, and impact exact agreement.

Table 3: Verification results for different comment inputs. EQ MAE is the mean absolute error of evidence quality scores.

Comment input	N	Verdict exact agreement ↑	Macro-F1 ↑	EQ MAE ↓	Impact exact agreement ↑
Original comment	200	45.50%	29.20%	0.86	42.50%
Neutralized concern	200	48.00%	37.00%	0.90	39.50%

As shown in Table 3, with paper evidence held constant, the neutralized concern achieves higher verdict classification performance. Compared with the original comment, verdict exact agreement increases from 45.50% to 48.00%, while Macro-F1 increases from 29.20% to 37.00%. These are gains of 2.50 and 7.80 percentage points, respectively. The original comment performs better on the other evaluation dimensions. Its EQ MAE is 0.86, lower than the 0.90 obtained with the neutralized concern, and its impact exact agreement is 42.50%, which is 3.00 percentage points higher. These results show that, on the current test set, neutralization improves verdict classification performance for evidence support, but does not improve evidence quality or impact assessment at the same time.

These results show that the original comment and the neutralized concern serve different roles in the verification task. The original comment describes the issue being checked while retaining the commenter's discussion of the background, evidence links, and possible consequences. This context may help the model assess impact. However, the original comment also contains evaluative wording, tone, and a judgment framework. The model may therefore be influenced by the commenter's stance and may fail to distinguish the commenter's judgment from the conclusion actually supported by the paper evidence. The neutralized concern focuses the input on the claim being checked and may reduce the interference of such wording in verdict judgments. Retaining the original comment therefore does not necessarily lead to more accurate verification. Its contextual value and potential wording bias should be considered separately, and a more informative input does not necessarily produce better overall performance.

4.5 EVIDENCE INPUT ABLATION

To separate visual information, paper evidence, and evidence and concern matching, we compare four conditions on the same 200 cases. Table 4 reports the results. `neutral-multimodal` supplies the neutralized concern with matched paper text and images. `text` removes the images but keeps the matched paper text. `claim_only` supplies only the concern. And `mismatched` supplies text and images from another paper. The Student model, instructions, output format, and decoding configuration are the same in every condition.

Table 4: Evidence input ablation. Each condition contains 200 cases.

Condition	Verdict exact agreement ↑	Macro-F1 ↑	EQ MAE ↓	Impact exact agreement ↑
`neutral-multimodal`	48.00%	37.00%	0.89	39.50%
`text`	42.50%	29.10%	1.23	23.50%
`claim_only`	38.50%	13.90%	1.69	11.00%
`mismatched`	25.50%	11.90%	1.98	8.50%

As shown in Table 4,without paper evidence, claim only verdict exact agreement, Macro-F1, and impact exact agreement fall to 38.50%, 13.90%, and 11.00%, while EQ MAE rises to 1.69. The full multimodal condition reaches 48.00%, 37.00%, 39.50%, and 0.89 on the same four measures. This shows that the concern text alone is not enough to reproduce the expert judgment based on paper materials. Against text only input, multimodal input raises verdict exact agreement and Macro-F1 by 5.50 and 7.90 points, lowers EQ MAE by 0.34, and raises impact exact agreement by 16.00 points. Compared with `claim_only`, `mismatched` has lower agreement scores and a higher EQ MAE. Correctly matched paper evidence is necessary, and images contribute information beyond text. The ablation also shows that unrelated material cannot replace the evidence linked to the concern: adding more input does not compensate for a broken evidence concern correspondence. A useful PubPeer verification assistant therefore needs both textual and visual paper evidence and must preserve their link to the concern.

5 DISCUSSION

The comment wording experiment shows that phrasing affects the cues used for different judgments. Neutralized concerns keep attention on the claim being checked and reduce interference from the commenter's stance, evaluative wording, and complex argument structure. This matches their higher verdict exact agreement and Macro-F1. Complete comments retain background, evidence links, and possible consequences, which may provide more context for evidence quality and impact assessment. This is consistent with their stronger scores on those metrics.

The results suggest that comment information depends on both its amount and its organization relative to the evaluation target. Evaluative wording and a commenter's framing may help a model identify the consequences under discussion, but they may also draw it toward rhetorical strength instead of paper evidence. Neutralization changes wording, organization, and compression at the same time, so the observed differences cannot be assigned to tone alone. The small EQ MAE difference likewise means that these explanations are inferences from the current test results.

6 CONCLUSION

We studied how to verify scientific concerns raised in PubPeer comments. We built PubPeerBench-1000 with 1,000 cases and trained Student Geng's WorkBuddy-27B from Qwen3.6-27B with LoRA supervised fine tuning. The task links each concern to paper text, images, and related materials and asks the model to assess evidence support, evidence quality, and impact. Student Geng's WorkBuddy-27B improves verdict exact agreement, Macro-F1, impact exact agreement, and EQ MAE over the untuned base model, and it outperforms all leading general purpose models included in the comparison, including GPT-6 Astra, on the reported metrics. These results support its use as an assistant for checking scientific concerns in PubPeer comments.

AI USE STATEMENT

During the preparation of this manuscript, we utilized a large language model to aid and polish the writing. The LLM served as a general-purpose assistant for improving grammar, clarity, and phrasing. All content was reviewed and edited by the authors.

ETHICS STATEMENT

This work uses publicly available PubPeer comments and related scientific materials to support systematic and reproducible verification of scientific concerns. We preserve source attribution and follow applicable permissions and usage constraints in dataset construction and release.

REPRODUCIBILITY STATEMENT

The main paper describes the case construction, annotation protocol, data split, model training, and evaluation metrics. Appendix A specifies the case schema and fields, Appendix B defines the annotation labels and adjudication protocol, Appendix C provides the data split and evaluation details, Appendix D specifies the experimental input conditions and evaluation protocol, Appendix E reports the model implementation and training settings, and Appendix F further documents the expert annotation process and agreement statistics.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Alibaba Cloud. Alibaba unveils qwen3.8-max: Its largest and most capable flagship model to date, 2026. URL https://www.alibabacloud.com/en/press-room/alibaba-unveils-qwen3-8-max.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information, 2021. URL https://arxiv.org/abs/2106.05707.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7352–7364, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.656. URL https://aclanthology.org/2020.acl-main.656/.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

Elisabeth M. Bik, Arturo Casadevall, and Ferric C. Fang. The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, 7(3), 2016. ISSN 2150-7511. doi: 10.1128/mbio.00809-16. URL http://dx.doi.org/10.1128/mBio.00809-16.

Sergio Bravo-Solorio and Asoke K. Nandi. Automated detection and localisation of duplicated regions affected by reflection, rotation and scaling in image forensics. *Signal Processing*, 91(8): 1759–1770, 2011. ISSN 0165-1684. doi: 10.1016/j.sigpro.2011.01.022. URL http://dx.doi.org/10.1016/j.sigpro.2011.01.022.

Enrico M. Bucci. Automatic detection of image manipulations in the biomedical literature. *Cell Death & Disease*, 9(3), 2018. ISSN 2041-4889. doi: 10.1038/s41419-018-0430-3. URL http://dx.doi.org/10.1038/s41419-018-0430-3.

Jennifer A. Byrne, Anna Abalkina, Olufolake Akinduro-Aje, Jana Christopher, Sarah E. Eaton, Nitin Joshi, Ulf Scheffler, Nick H. Wise, and Jennifer Wright. A call for research to address the threat of paper mills. *PLOS Biology*, 22(11):e3002931, 2024. ISSN 1545-7885. doi: 10.1371/journal.pbio.3002931. URL http://dx.doi.org/10.1371/journal.pbio.3002931.

Minal M. Caron, Carolyn T. Lye, Barbara E. Bierer, and Mark Barnes. The pubpeer conundrum: Administrative challenges in research misconduct proceedings. *Accountability in Research*, 32 (8):1369–1387, 2024. ISSN 1545-5815. doi: 10.1080/08989621.2024.2390007. URL http://dx.doi.org/10.1080/08989621.2024.2390007.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2022. URL https://arxiv.org/abs/2209.06794.

Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou. An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on Information Forensics and Security*, 7(6):1841–1854, 2012. ISSN 1556-6021. doi: 10.1109/tifs.2012.2218597. URL http://dx.doi.org/10.1109/TIFS.2012.2218597.

Marcelo Cicconet, H Elliott, David L Richmond, D Wainstock, and M Walsh. Image forensics: Detecting duplication of scientific images with manipulation-invariant image similarity. *arXiv preprint arXiv:1802.06515*, 2018.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Daniele Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5):e5738, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0005738. URL http://dx.doi.org/10.1371/journal.pone.0005738.

Ferric C. Fang, R. Grant Steen, and Arturo Casadevall. Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, 109(42): 17028–17033, 2012. ISSN 1091-6490. doi: 10.1073/pnas.1212247109. URL http://dx.doi.org/10.1073/pnas.1212247109.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020. URL https://arxiv.org/abs/2002.08909.

Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. Transforensics: Image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15055–15064, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hao_TransForensics_Image_Forgery_Localization_With_Dense_Self-Attention_ICCV_2021_paper.html.

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The extent and consequences of p-hacking in science. *PLOS Biology*, 13(3):e1002106, 2015. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002106. URL http://dx.doi.org/10.1371/journal.pbio.1002106.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4651–4664, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. HoVer: A dataset for many-hop fact extraction and claim verification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3441–3460, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.309. URL https://aclanthology.org/2020.findings-emnlp.309/.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. URL https://aclanthology.org/2020.emnlp-main.550/.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL http://dx.doi.org/10.1145/3397271.3401075.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5430–5443, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.474. URL https://aclanthology.org/2020.coling-main.474/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings*

of the 40th International Conference on Machine Learning, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Moonshot AI. Kimi k3 technical report, 2026. URL https://github.com/MoonshotAI/Kimi-K3/blob/main/k3_tech_report.pdf.

Dennis Normile. Science sleuth finds rare acclaim in china. *Science*, 392(6803):1102–1103, 2026. doi: 10.1126/science.aej5918. URL https://doi.org/10.1126/science.aej5918.

Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226, 2015. ISSN 1554-3528. doi: 10.3758/s13428-015-0664-2. URL http://dx.doi.org/10.3758/s13428-015-0664-2.

OpenAI. Gpt-6 astra: A new generation of intelligence, 2026. URL https://openai.com/index/gpt-6-astra/.

José Luis Ortega. Classification and analysis of pubpeer comments: How a web journal club is used. *Journal of the Association for Information Science and Technology*, 73(5):655–670, 2022. doi: https://doi.org/10.1002/asi.24568. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24568.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL https://arxiv.org/abs/2306.14824.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Ekraam Sabir, Soumyaroop Nandi, Wael Abd-Almageed, and Prem Natarajan. Biofors: A large biomedical image forensics dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10963–10973, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Sabir_BioFors_A_Large_Biomedical_Image_Forensics_Dataset_ICCV_2021_paper.html.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. Evidence-based fact-checking of health-related claims. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3499–3512, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.297. URL https://aclanthology.org/2021.findings-emnlp.297/.

R Grant Steen. Retractions in the scientific literature: is the incidence of research fraud increasing? *Journal of Medical Ethics*, 37(4):249–253, 2010. ISSN 1473-4257. doi: 10.1136/jme.2010.040923. URL http://dx.doi.org/10.1136/jme.2010.040923.

Fiona Townsend. Post-publication peer review: Pubpeer. *Editors' Bulletin*, 9(3):45–46, 2013. ISSN 1752-1742. doi: 10.1080/17521742.2013.865333. URL http://dx.doi.org/10.1080/17521742.2013.865333.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-open: Towards open-domain scientific claim verification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347/.

Ziyue Xiang and Daniel E. Acuna. Scientific image tampering detection based on noise inconsistencies: A method and datasets, 2020. URL https://arxiv.org/abs/2001.07799.

Xiaoying You. Student geng ignites research-integrity scandal in china after calling out senior academics. *Nature*, 655:267–269, 2026. doi: 10.1038/d41586-026-01902-0. URL https://www.nature.com/articles/d41586-026-01902-0.

Jie Zhao and Weifeng Zhao. Passive forensics for region duplication image forgery based on harris feature points and local binary patterns. *Mathematical Problems in Engineering*, 2013:1–12, 2013. ISSN 1563-5147. doi: 10.1155/2013/619564. URL http://dx.doi.org/10.1155/2013/619564.

A Dataset Schema and Case Fields

Each benchmark case is represented as a versioned record that preserves the link between the public concern, the target paper, the retrieved evidence, and the expert judgment. Table 5 summarizes the fields used in the released case representation. The complete paper JSON is retained for archival traceability, while the evidence JSON bundle contains the material selected for a particular claim.

Table 5: Case fields in PubPeerBench-1000.

Field	Description
`case_id`	Stable identifier used to connect the case to its source records.
Paper identifiers	Paper title, DOI, publication URL, and version or association information.
Comment record	Original PubPeer comment, comment number, commenter status, and attachments.
`claim`	Neutralized concern that retains checkable factual claims and their scope.
`issue_types`	Multi label description of the independently checkable issue types.
`evidence_scope`	Types of material needed to assess the concern, such as text, figures, methods, supplements, or source data.
Paper JSON	Parsed manuscript text, figures, tables, captions, methods, supplementary materials, source URLs, parsing status, and hashes.
Evidence JSON bundle	Claim linked text blocks, images, tables, locations, retrieval status, and a short evidence summary.
Expert labels	Final verdict, evidence quality, impact, and rationale after adjudication.

B Annotation Guidelines

The annotation protocol separates evidence support from evidence quality and potential impact. Annotators first identify the key claims in the comment and inspect the linked paper materials. They then assign the fields below. A missing source is recorded as `insufficient_evidence` when it prevents a judgment. Missing support is not automatically treated as evidence against the concern.

Table 6: Expert annotation labels and operational definitions.

Field	Labels and definitions
Verdict	`supported`: the available materials support the key claims. `partially_supported`: some key subclaims are supported. `unsupported`: the materials provide counter evidence or resolve the concern. `insufficient_evidence`: necessary materials are missing or cannot support a decision.
Evidence quality	Ordered score from 0 to 4. The score increases with relevance, locatability, reproducibility, and consistency across available sources.
Impact	`none`, `minor`, `moderate`, `major`, or `unknown`, describing the possible effect on the paper's results or interpretation if the concern is true.
Rationale	One to three sentences describing the observed evidence, its connection to the concern, and material limitations or alternative explanations.

Two experts annotate each case independently. A third expert resolves disagreements and produces the final expert labels. The independent annotations and adjudication rationale are retained for quality control, while all reported metrics use the final expert labels.

C Data Splits and Evaluation Details

PubPeerBench-1000 is divided into 700 training cases, 100 validation cases, and 200 independent test cases. Paper association groups are formed from DOI information, paper version relations, cross paper links, and near duplicate signals. No association group crosses a split. This procedure reduces the chance that text, images, or evidence from the same paper appears in both training and test data.

The reported metrics use the same 200 test cases for every model and input condition. Verdict is evaluated with exact agreement and Macro-F1. Impact is evaluated with exact agreement, with `unknown` treated as a separate state. Evidence quality is evaluated with EQ MAE, the mean absolute difference between the predicted and expert scores on the numeric encoding from 0 to 4. The rationale is retained as a structured output but is not assigned a numerical score in the current experiments.

D Input Conditions and Reproducibility

The standard model input contains the neutralized concern, matched paper text, and matched image evidence. The wording experiment replaces the neutralized concern with the complete original comment while keeping model weights, instructions, paper evidence, and decoding configuration fixed. The evidence ablation compares four conditions: `neutral-multimodal`, `text`, `claim_only`, and `mismatched`. The first uses matched text and images, the second removes images, the third removes paper evidence, and the fourth supplies evidence from another paper.

Student Geng's WorkBuddy-27B starts from Qwen3.6-27B and is trained with LoRA supervised fine tuning. Training uses BF16. The base model, vision encoder, and multimodal projection remain frozen, while LoRA adapters are updated in the attention and MLP layers of the language module. The same output schema is used for training and evaluation: `verdict`, `evidence_quality`, `impact`, and `rationale`.

E Model Implementation Details

This section records the implementation settings for the 27B model used in the reported experiments. The base model is Qwen3.6-27B. We use LoRA supervised fine tuning in BF16 without 4-bit or 8-bit quantization. The base model weights, vision encoder, and multimodal projection are frozen. Trainable adapters with rank 16, alpha 32, and dropout 0.05 are inserted into the language model's attention projections q, k, v, and o, and its MLP gate, up, and down projections. The optimizer is AdamW with weight decay 0.01 and a learning rate of 5×10^{-5}. We use cosine decay with a 5% warmup, corresponding to 7 warmup updates.

Training uses one example per device step, gradient accumulation over 8 steps, and two H200 GPUs with distributed data parallelism, giving a global batch size of 16. The fixed training budget is 3 epochs and 132 optimizer updates. The loss is computed only on the target JSON response and its end token. There is no patience based early stopping. We run seeds 42, 43, and 44. At the end of each epoch, we compute the mean negative log likelihood on answer tokens in the 100 case validation split and select the checkpoint with the lowest value. The test split is never used for parameter updates or checkpoint selection. All three seeds select epoch 3 at optimizer step 132. The reported main table uses the seed 42 checkpoint, named `27b_s42/epoch_3`. Its training loop took 124.8 minutes from model loading through validation and checkpoint saving.

The model context limit is 65,536 tokens, with up to 3,072 tokens reserved for generation. The largest constructed 27B input is 36,043 tokens, and the preprocessing check found no over length case. The training code raises an error rather than silently truncating an over length input. The input bundle contains the preselected paper pages associated with the case, rather than every page of the paper, all supplementary material, or every comment attachment. Each bundle contains between 0 and 11 pages. Pages are rendered at approximately 1,900 pixels across while preserving aspect ratio and saved as JPEG with quality 92. The Qwen processor applies dynamic resizing with a pixel budget from 65,536 to 2,097,152 pixels.

F Description of the Expert Annotation Process

Two experts annotate each case independently. Before adjudication, agreement between the two experts is measured on all 1,000 cases. The resulting Cohen's kappa values are 0.89 for verdict and 0.92 for impact. Evidence quality is recorded on an ordinal scale, so weighted Cohen's kappa is used, with a value of 0.81. These agreement statistics are computed from the two independent annotations before adjudication and are separate from the final expert labels. A third expert resolves

disagreements and produces the final expert labels. The independent annotations and adjudication rationale are retained for quality control, while all reported metrics use the final expert labels. Relative to the first expert's annotations, the final expert labels differ in 5.20% of `verdict` cases and 3.50% of `impact` cases. The weighted Cohen's kappa between the final expert evidence quality scores and the first expert's scores is 0.59.