
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247282/

Version: Published Version

Proceedings Paper:
James, J., Xiao, C., Li, Y. et al. (2026) RIGOURATE: Quantifying scientific exaggeration with evidence-aligned claim evaluation. In: Liakata, M., Moreira, V., Zhang, J. and Jurgens, D., (eds.) Findings of the Association for Computational Linguistics: ACL 2026. The 64th Annual Meeting of the Association for Computational Linguistics, 02-07 Jul 2026, San Diego, California, United States. ACL Anthology. Association for Computational Linguistics, pp. 34022-34043. ISBN: 9798891763951. ISSN: 0736-587X.

https://doi.org/10.18653/v1/2026.findings-acl.1699

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.


UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

RIGOURATE: Quantifying Scientific Exaggeration with Evidence-Aligned Claim Evaluation

Joseph James[1], **Chenghao Xiao**[2], **Yucheng Li**[3]
Nafise Sadat Moosavi[1], **Chenghua Lin**[4*]
[1]The University of Sheffield, UK [2] Durham University, UK
[3] University of Surrey, UK [4] The University of Manchester, UK
jhfjames1@sheffield.ac.uk chenghua.lin@manchester.ac.uk

Abstract

Scientific rigour tends to be sidelined in favour of bold statements, leading authors to overstate claims beyond what their results support. We present RIGOURATE, a two-stage multimodal framework that retrieves supporting evidence from a paper's body and assigns each claim an overstatement score. The framework consists of a dataset of over 10K claim–evidence sets from ICLR and NeurIPS papers, annotated using eight LLMs, with overstatement scores calibrated using peer-review comments and validated through human evaluation. It employs a fine-tuned reranker for evidence retrieval and a fine-tuned model to predict overstatement scores with justification. Compared to strong baselines, RIGOURATE enables improved evidence retrieval and overstatement detection. Overall, our work operationalises evidential proportionality and supports clearer, more transparent scientific communication. We release both our models[1] and code[2].

1 Introduction

Effective scientific writing and reviewing demands not only a clear presentation of novel ideas but also a rigorous grounding of findings that support them. In many research papers, authors are incentivised to write abstracts and introductions that capture readers' attention by showcasing their contributions in an eye-catching manner (Bavdekar, 2015; Rahman et al., 2017; Kawase, 2018; Hyland and Jiang, 2021; Intemann, 2022; Stavrova et al., 2025). However, when such claims are exaggerated they can mislead the reader if robust evidence is not provided in later sections. In a rapidly evolving field such as Machine Learning (Pineau et al., 2021), this dynamic has fostered a "Publish or Perish" environment (De Rond and Miller, 2005; Rawat and Meena, 2014). The pressure to publish quickly can lead authors to prioritise speed over scientific rigour, thus accumulating "scientific debt" (Nityasya et al., 2023), where well-grounded, reproducible work is sidelined in favour of rapid publication, diminishing the quality of scientific research.

We focus on the phenomenon of overstatement: rhetorical exaggeration in which the wording of a claim amplifies its strength beyond what the paper's evidence supports. We cast this as a distinct aspect of fact-checking, namely quantifying degrees of overstatement, where claims are rarely outright false but often exceed what the available evidence warrants. Rather than making binary true/false judgments, we assess whether the strength and scope of a claim are proportionally grounded in the paper's own methods and results. A claim is overstated when it presents an inflated representation that is not adequately supported by the paper, for example due to limited evidence or unjustified generalisation. This perspective highlights how linguistic choices and rhetorical emphasis can shape readers' perceptions of contribution even before they encounter the technical sections, thereby influencing how the communicative rigour of a claim and its alignment with the supporting evidence are perceived.

Building on this motivation, we define the task of *intra-paper overstatement detection*, which evaluates whether claims in the abstract and introduction are proportionally supported by evidence presented in the remainder of the paper. We distinguish overstatement detection from adjacent research areas referenced throughout: fact-checking and claim verification evaluate factual accuracy against external knowledge sources, while reproducibility research concerns whether findings can be independently replicated. Our task is strictly intra-paper, assessing whether the strength of a claim is proportionate to the evidence presented within the same

* Corresponding author
[1] https://hf.co/collections/Joeyhfj/rigourate
[2] https://github.com/Rexathon/RIGOURATE

Findings of the Association for Computational Linguistics: ACL 2026, pages 34022–34043
July 2-7, 2026 ©2026 Association for Computational Linguistics

paper.

To address this problem, we introduce RIGOURATE, a multimodal, review-informed, automated framework that tackles two tasks: (i) evidence retrieval and (ii) overstatement detection with reasoning. We collect papers and reviews from ICLR and NeurIPS hosted on OpenReview, a venue with openly accessible reviews covering diverse NLP and Machine-Learning topics. We employ a panel of large language model (LLM) annotators to identify claims that are the authors' own statements. Then we extracted potential evidence spanning text, figures, and tables, classifying each passage as either *relevant* (directly related to the claim) or *irrelevant*. Since the core issue is the degree of support rather than direct refutation, we focus on how thoroughly the available evidence grounds each claim. We assign every claim a continuous overstatement score ranging from 0 to 1. The LLM annotators first generate a score for each claim–evidence set, and then we incorporate each review comment as additional context. This design reduces sensitivity to individual LLM and reviewer perspectives, with annotation quality validated through targeted human evaluation.

We use RIGOURATE to assess the feasibility and validity of intra-paper overstatement detection. The evaluation consists of two components: (i) evidence retrieval, which tests whether models can reliably identify passages, figures, and tables that support a given claim, and (ii) overstatement detection, which assesses whether models can estimate the degree to which a claim is proportionally supported by that evidence. We adopt a range of existing state-of-the-art reranker and multimodal models to this setting. The results show that fine-tuning enables these models to meaningfully learn the task signal, yielding consistent improvements across retrieval metrics and more accurate overstatement scoring compared to base zero-shot models. These findings indicate that claim–evidence alignment within a paper is a learnable and evaluable problem, and that RIGOURATE provides a practical framework for computationally assessing rhetorical overstatement via evidential proportionality, supporting automated, evidence-based evaluation of research claims and more transparent scientific communication. To summarise, our main contributions are three-fold:

- A review-informed framework that automatically extracts and scores multimodal claim–evidence sets within research papers.
- We show that intra-paper claim–evidence alignment is a learnable task, with fine-tuning consistently improving performance on evidence retrieval and overstatement scoring.
- A case study demonstrating that our framework is capable of identifying rhetorical amplification, where overstatement often stems from missing substantive detail and surface-level phrasing.

2 Related Work

2.1 Scientific Rigour

The integrity of scientific research is increasingly challenged by issues of reproducibility, alongside broader concerns about how claims are framed and interpreted. While reproducibility and reporting practices have been widely studied (ICML and ICLR, 2019; ML Reproducibility Challenge, 2025), comparatively less attention has been paid to the alignment between the strength of scientific claims and the evidence presented to support them. The framing of a paper's contributions can strongly shape perceptions of its value and rigour, allowing novel but non-replicable claims to gain prominence when independent verification is limited (Salager-Meyer, 1994; Ferrari Dacrema et al., 2019; Gustafson and Rice, 2020; Serra-Garcia and Gneezy, 2021; James et al., 2024). Although such non-replicability is not necessarily caused by exaggerated framing, it reduces opportunities for empirical scrutiny, allowing weakly supported claims to persist and thereby undermining research credibility (Raghupathi et al., 2022).

The NeurIPS and ARR checklist have addressed several challenges, including insufficient exploration of variables, poor documentation, and a lack of reporting of crucial details needed to replicate results (Pineau et al., 2021; ARR, 2025). While the checklist approach has improved the quality of research, it requires reviewers to manually validate whether the claims made are backed up by sufficient evidence. By detecting overstated claims automatically, our work aims to support authors in presenting their work accurately and to provide a framework for evaluating the rigour in research claims. Evidence that this remains challenging is provided by peer review analyses showing that even highly rated papers often receive requests for additional experiments, indicating unresolved gaps between claims and supporting evidence (Wang

et al., 2023). This motivates the need for systematic methods to assess claim–evidence alignment.

2.2 Scientific Claim Verification

The increasing number of publications requires the development of automated methods for verifying research claims. Scientific fact verification, which aims to assess the accuracy of scientific statements, often relies on *external knowledge* to support or refute claims (Wadden et al., 2022; Vladika and Matthes, 2023; Dmonte et al., 2024). More broadly, fact-checking encompasses multiple dimensions beyond binary factual correctness, including understatement, exaggeration, and contradiction (Kao and Yen, 2024). However, existing approaches rarely assess whether the strength of a claim is proportionate to the evidence within the same paper.

Recent work has highlighted the importance of grounding claims in paper-internal evidence, with Chan et al. (2024) collecting claims linked to lab notes, figures, and methodological details to enable more context-aware claim evaluation. Schlichtkrull et al. (2023) examine how automated fact-checking methods are framed and motivated in highly cited NLP papers, particularly in introductions, showing that claims about verification systems are often underspecified with respect to their intended use and scope, highlights the need for clearer articulation of what different forms of verification are designed to assess.

In contrast to prior work on scientific claim verification, our approach targets a distinct aspect of fact-checking: assessing evidence proportionality within a single paper. Rather than determining factual correctness, we evaluate whether the strength and scope of a claim are justified by the paper's own methods and results. We introduce a granular overstatement score to capture degrees of exaggeration, as claims are rarely contradicted by internal evidence but are often phrased more strongly than that evidence warrants. This positions our work as complementary to existing fact-checking efforts, focusing on scientific rigour and clarity rather than external verification.

2.3 Automatic Peer Reviewing

Rising workloads placed on reviewers makes automating aspects of the peer-review process increasingly important (Staudinger et al., 2024; Eger et al., 2025). However, LLMs often lack the domain knowledge required to critique methodological details (Du et al., 2024). Benchmarks such as AAAR-1.0 (Lou et al., 2025) look into identifying paper weaknesses and reliability of reviews, and show that models can fall short in detecting subtle weaknesses.

Building on prior work showing that incorporating peer-review comments improves LLM-based evaluation accuracy (Zhou et al., 2024) and that LLMs attend selectively to different aspects of scientific feedback (Liang et al., 2024), we derive review-informed overstatement scores during annotation. This design is motivated by the observation that evaluating soundness, comparisons, and substantive claims is knowledge-intensive when relying on paper content alone. Conditioning models on reviews provides access to expert-written critiques of evidential sufficiency and overgeneralisation, yielding more consistent and calibrated overstatement judgments, consistent with prior work demonstrating the value of review text for modelling peer-review outcomes (Bharti et al., 2024).

Recent ML conferences position LLMs as lightweight assistants rather than replacements in peer reviewing, supporting reproducibility checks and review quality without influencing editorial decisions (NeurIPS, 2025; ICLR, 2025; AAAI, 2025). In contrast to these systems, which focus on aiding the review process itself, our work leverages peer-review signals to assess whether authors' claims are proportionally supported by the evidence presented in their papers.

3 Data Processing Framework

Task Definition. We define two tasks for detecting overstatements in claims, focusing on claims extracted from the abstract and introduction of scientific papers: (i) *Evidence Retrieval*: Given a claim, retrieve all relevant evidence that directly supports the claim. (ii) *Overstatement detection*: Given a claim and its corresponding evidence, assign a continuous score indicating the degree to which the claim's wording exceeds what the evidence supports, accompanied by a justification. A claim is *overstated* when it makes assertions not justified by the paper's evidence (e.g., limited experiments, lack of methodological detail, etc); *partially overstated* when some components are supported but others extend beyond what the evidence warrants; and *well-stated* when the claim is fully grounded in the paper's methods, results, and reasoning without exaggeration. Examples of each claim type are provided in the case study in Table 5. Our goal

is therefore to operationalise epistemic overstatement: the misalignment between the strength of a claim and the evidence provided within the same paper, rather than rhetorical intensity per se. Disciplinary conventions that favour emphatic language are not treated as problematic; only claims whose strength exceeds what the paper's own evidence warrants are considered overstated. Overstatement may stem from linguistic exaggeration (e.g., *prove*, *always*), evidential insufficiency, or both; claims that appropriately hedge uncertainty (e.g., *suggest*, *may*) are one example of well-stated claims, provided their strength remains proportionate to the evidence.

3.1 Data Preparation

We collected papers and associated reviews from OpenReview, focusing on NeurIPS and ICLR submissions. In addition to using the OpenReview API[3], we incorporated previously collected ICLR datasets (Wang et al., 2020; Yuan et al., 2022; Li et al., 2023; Wang et al., 2023).[4] To mitigate reviewer subjectivity and ensure consistency, we restricted our dataset to papers for which all reviewers assigned identical overall scores. Prior work has shown that reviewer disagreement is common and can substantially affect review outcomes, with repeated evaluations often leading to different accept/reject decisions (Beygelzimer et al., 2023). Review quality has also been shown to vary due to bias and miscalibration across evaluators (Goldberg et al., 2025). Focusing on high-agreement subsets is therefore a common strategy for constructing more reliable peer-review benchmarks (Staudinger et al., 2024; Peng et al., 2025).

We processed PDFs using SciPDF[5] for text extraction and PDFFigures2 (Clark and Divvala, 2016) for tables and figures, segmenting papers into paragraphs, figures, and tables. Due to PDF parsing failures, some papers were excluded, resulting in 659 ICLR and 213 NeurIPS papers. Reviewer scores are well distributed across the dataset (see Appendix E).

To improve evidence retrieval precision, we segment papers into smaller textual units, allowing models to operate over sentences rather than long contexts, which has been shown to improve retrieval accuracy for LLM-based reviewing tasks (Zhou et al., 2024). Claims are extracted from the abstract and introduction, while supporting evidence is drawn from the main body, including text, tables, and figures with captions. We employ a multi-LLM annotation framework consisting of three text-only (SEED-OSS-36B-INSTRUCT,GPT-OSS-120B, and DEEPSEEK-R1-DISTILL-QWEN-32B) and five vision-language models (VLMs) (GEMMA-3-27B-IT, APRIEL-1.5-15B-THINKER, KIMI-VL-A3B-INSTRUCT, MINICPM-V-4_5, and QWEN3-VL-30B-A3B-INSTRUCT), each independently annotating claims and evidence (Team, 2025; Agarwal et al., 2025; DeepSeek-AI, 2025a; Kamath et al., 2025; Radhakrishna et al., 2025; Team et al., 2025; Yu et al., 2025; Yao et al., 2025; Qwen, 2025). Full model names provided in Appendix A. Final annotations are determined by majority vote, reducing model-specific bias and improving consistency across annotators (Pavlovic and Poesio, 2024; Liu et al., 2024; Tseng et al., 2025; Yuan et al., 2025).

3.2 Author's Own Statement Extraction

To extract claims, we define a *claim* as any sentence that clearly presents an original claim, finding, or result that is central to the paper's contributions. This definition excludes sentences that offer only background, contextual information, or references of prior work. We first split the abstract and introduction into individual sentences using WTPsplit (Frohmann et al., 2024), as it consistently outperformed other sentence segmentation baselines, particularly on scientific text. For each sentence, we provided the complete abstract and introduction as context and then prompted our LLMs to classify the sentence (prompt in Table 14 in Appendix G). In total **10,641** claims were extracted as author's own statements. To assess the robustness of the annotation process, we compute Krippendorff's α between the full-panel majority vote and a leave-one-out majority vote for each annotator model. Specifically, we recompute the majority label after excluding that model and measure agreement with the full-panel consensus. High agreement indicates that no single model disproportionately influences the final annotation and that the consensus labels are robust to the removal of individual annotators. Across all models, this analysis shows near-perfect agreement, suggesting stable aggregation rather than dominance by any single model (full break-

[3] https://github.com/openreview/openreview-py

[4] https://github.com/hughplay/ICLR2024-OpenReviewData

[5] https://github.com/titipata/scipdf_parser

down shown in Appendix A).[6]

3.3 Evidence Retrieval

We segment each paper's main body into sentences using WTPsplit and assign sentence IDs so LLMs can reference evidence by index rather than copying text, reducing hallucinations. For each claim, we prompt the LLM annotators to select supporting evidence from the paper body, prioritising results, analysis, conclusions, and other directly relevant context while excluding irrelevant or purely paraphrased content (prompt in Appendix G, Table 15). Selected adjacent sentences are merged into coherent passages to preserve local context. For tables and figures, LLM annotators assess whether the visual content and caption support the claim; tables are treated as images, with their content evaluated by VLMs using the table image and caption. Following findings that retrieval quality degrades with long contexts (Modarressi et al., 2025), we keep annotation contexts under $\sim$1K tokens. We further compute Krippendorff's α between the full-panel majority vote and a leave-one-out majority vote for each model. Overall, agreement across models ranges from substantial to almost perfect, indicating that the aggregated annotations are stable to the removal of individual annotators (full breakdown shown in Appendix A).[7] Manual inspection shows that the majority of disagreements are driven by span-level variation: annotators typically agree on the core supporting evidence but differ in how much surrounding context they include. Additional variability arises from the presence of multiple valid supporting passages, with some models retrieving only a subset of the relevant evidence, while others also focus on weaker or more indirect supporting passages for the same claim.

3.4 Overstatement Annotation

We adopt a review-informed LLM annotation strategy for scoring claim overstatement, motivated by prior work showing that LLM-based evaluation benefits from peer-review context (Zhou et al., 2024; Liang et al., 2024). For each claim-evidence set, the same LLM assigns a continuous overstatement score in the range $[0, 1]$, where 0 denotes a well-stated claim and 1 denotes a clearly overstated claim. We obtain multiple scores for each claim-evidence set under different annotation contexts. In a paper-only setting, the model relies solely on the paper content to assess overstatement. In review-informed settings, the same model is additionally conditioned on individual peer-review comments and produces a separate score for each review.[8] Conditioning on reviewer feedback exposes the model to expert-written critiques of evidential sufficiency and overgeneralisation, yielding judgments that more closely reflect reviewer reasoning. Because overstatement is inherently graded and admits legitimate disagreement, we retain all individual scores rather than aggregating them via majority voting. This produces a denser supervision signal that captures variability across models and annotation contexts. For the validation and test splits, we compute the mean score across annotations for each claim-evidence set and use this average as a soft label.

Split	Train	Dev	Test	Total
Paper IDs	536	259	77	872
Claims	6,449	3,056	1,063	10,641
Evidence	429,519	190,414	62,038	681,971
Scores	159,930	3,056	1,063	164,049

Table 1: Dataset statistics. Evidence includes both supporting and not-supporting items, full breakdown shown in Appendix B.

Quality Control. We assess the reliability of the automatically assigned overstatement scores via a human validation study with two PhD-level evaluators in Computer Science and Machine Learning. Each evaluator independently rated 30 claim-evidence sets sampled across the full score range, including both textual and visual evidence. Ratings were provided on a five-point ordinal scale (1–5), where 1 denotes a well-stated claim fully supported by evidence and 5 denotes a clearly overstated claim. Evaluators were provided with a definition of overstatement consistent with our task formulation: a claim is overstated when its wording exceeds what the paper's own evidence supports due to limited experiments, insufficient methodological detail, or unjustified generalisation, irrespective of factual correctness or novelty relative to prior work.

Model predictions were discretised into ordinal bins (equal-width intervals corresponding to the 1–5 human scale) and compared against hu-

[6] Manual check of 200 randomly sampled sentences showed that the LLMs correctly identified over 98% of cases.

[7] Manual check of 100 randomly sampled pairs showed that the LLMs correctly identified over 91% of cases.

[8] Peer reviews are provided as a single contextual unit, as they are written as holistic assessments whose reasoning spans multiple sentences, unlike localised evidence in the paper.

man ratings using Krippendorff's α (ordinal). The resulting agreement of **0.62** indicates substantial alignment between automated and human judgments (Wong et al., 2021). We additionally report the Intraclass Correlation Coefficient (ICC), a reliability measure for continuous ratings (Koo and Li, 2016). ICC(2,1) assesses absolute agreement assuming raters are a random sample, while ICC(3,1) assesses consistency assuming raters are fixed. Inter-rater reliability between the two human evaluators was also strong, with ICC(2,1) of **0.78** (95% confidence interval of [0.59, 0.89]) and ICC(3,1) of **0.80** (95% confidence interval of [0.62, 0.90]), both falling within the range typically interpreted as good reliability (Koo and Li, 2016). The similarity between ICC(2) and ICC(3) suggests that agreement was not driven by rater-specific bias. Descriptively, evaluators reached exact agreement on 60% of pairs, differed by one scale point on 36.7%, and diverged by more than one point on only 3.3%, indicating that disagreements were minor and infrequent.

Annotation Stability. Incorporating peer-review context introduces a small but systematic directional shift in overstatement scores. On average, review-informed scores increase relative to paper-only scores by an average of 0.028 (median 0.005), indicating that models become modestly more critical when exposed to reviewer feedback. This shift is asymmetric: 51.0% of scores increase after conditioning on reviews, compared to 32.6% that decrease, while 16.4% remain unchanged. Stratified analysis (Table 8 in Appendix) shows that reviews tend to raise scores for initially low or borderline claims, while slightly tempering scores for already high-overstatement claims, suggesting a calibration effect rather than uniform inflation. Despite these shifts, paper-only and review-informed scores remain strongly correlated (Pearson $r = 0.79$), showing that review context primarily refines existing judgments rather than overturning them, and increases the mean pairwise Pearson correlation across LLM annotators by 10.5%, reflecting greater consistency in their relative assessments. To assess potential annotator bias, we conducted a leave-one-model-out analysis over all claims. For each model, we recomputed the aggregated score without its annotations and compared the resulting distribution to the full baseline using Welch's t-test. Although several exclusions produced statistically significant differences ($p < 0.01$), all absolute shifts were small (MAD < 0.03), indicating that no single model disproportionately influences the final scores (see Appendix A, Table 7).

To prevent data leakage and assess cross-domain generalisation, we split the dataset by paper ID and exclude NeurIPS papers from the training set. Dataset statistics are summarised in Table 1.

4 Experimental Setup

Evidence Retrieval Setup. We evaluate the learnability of intra-paper evidence retrieval using supervision derived from RIGOURATE annotations. Only text information is utilised for the task, with the captions for the tables and figures in place of the visual inputs. We consider a diverse set of reranker models spanning bi-encoder, cross-encoder, and generative architectures, selected based on strong performance on the MTEB benchmark.[9] These include MiniLM (Reimers and Gurevych, 2019), BGE-Reranker (Chen et al., 2024), GTE-Reranker (Zhang et al., 2024), GritLM-7B (Muennighoff et al., 2025), and Qwen3-Reranker (Zhang et al., 2025). We additionally evaluate the E2Rank family, which combines dense retrieval with LLM-based relevance scoring (Liu et al., 2025). To assess whether the automatically constructed supervision generalises beyond the annotation process, we fine-tune the top 3 best performing model families. Full model specifications, training procedures, and prompting details are provided in Appendix B.

As claims may be supported by multiple evidence items, we report MAP to assess overall ranking quality, MRR (Voorhees et al., 1999) to measure how quickly relevant evidence is retrieved, Recall@k to evaluate coverage of supporting evidence, and NDCG@k (Järvelin and Kekäläinen, 2002) to reward placing the most informative evidence higher in the ranking.

Overstatement Detection Setup. We selected a range of state-of-the-art text-only and VLMs spanning multiple families and sizes. Specifically, we used DeepSeek (V3.2 and R1) (DeepSeek-AI, 2025b), and GLM-4.6 (Zeng et al., 2025) for text-only evaluation, and InternVL3.5 (38B and 30B-A3B) (Wang et al., 2025), Ovis2-34B (Lu et al., 2024), Qwen3-VL-32B (Qwen, 2025), GLM-4.5V (GLM-V et al., 2025), GPT-5-mini (low, high) (OpenAI, 2025) for multimodal analysis. We fine-tune several VLMs using our dataset to evaluate the role of visual evidence in the task. Specifically,

[9] https://huggingface.co/spaces/mteb/leaderboard

Model	MAP	MRR	R@5	R@10	R@20	N@5	N@10	N@20
Zero-shot models								
MiniLM-L6-v2	44.08*	66.12	10.37	20.46	39.72	71.48	71.66	71.82
BGE-Reranker-v2-m3	43.99	65.30	10.35	20.23	39.73	70.94	71.25	71.55
GTE-Reranker-base	43.99	65.80	10.47	20.40	39.43	71.54	71.62	71.84
GritLM-7B	43.15	66.17	10.16	19.47	38.22	70.83	71.39	71.36
Qwen3-Reranker-0.6B	44.19	66.73	10.54	20.71	39.77	72.07	72.05	72.14
Qwen3-Reranker-4B	44.92	68.74	10.93	21.02	40.71	73.37	73.13	73.12
Qwen3-Reranker-8B	45.88*	71.72	11.09	21.52	41.49	75.34	74.65	74.28
E2Rank-0.6B	47.57*	85.54	12.86	22.80	41.18	85.19	81.75	79.20
E2Rank-4B	47.41	86.06	12.67	22.68	41.07	85.16	81.63	79.13
E2Rank-8B	47.54	85.77	12.57	22.78	41.36	85.36	81.80	79.25
Fine-tuned (Top-3 models only)								
MiniLM-L6-v2	45.51 (+1.4)	74.62 (+8.5)	11.33 (+1.0)	21.37 (+0.9)	40.00 (+0.3)	77.41 (+5.9)	76.14 (+4.5)	75.02 (+3.2)
Qwen3-Reranker-8B	**54.19** (+8.3)	83.44 (+11.7)	**14.84** (+3.8)	**27.32** (+5.8)	**48.26** (+6.8)	85.19 (+9.9)	83.33 (+8.7)	**82.11** (+7.8)
E2Rank-0.6B	52.37 (+4.8)	**86.57** (+0.4)	14.69 (+1.2)	25.96 (+3.2)	46.34 (+5.1)	**86.75** (+1.6)	**83.95** (+2.2)	82.06 (+2.1)

Table 2: Task 1: Retrieval performance using MAP, MRR, Precision@5/10/20, and NDCG@5/10/20 (N). Fine-tuning for the top-3 rerankers (MiniLM-L6-v2, Qwen3-Reranker-8B, and E2Rank-0.6B), indicated by the * next to MAP scores. Green values in brackets indicate relative gains against the Base setting. **Bold** = best; underline = 2nd best.

we selected Qwen3-VL-8B-Instruct (Qwen, 2025), InternVL3.5-8B (Wang et al., 2025), and LLaVA-OV-1.5-8B-Instruct (An et al., 2025). See Appendix B for model specifications and fine-tuning details.

For evaluation, we use the concordance correlation coefficient (CCC) (Lawrence and Lin, 1989), which is well suited for overstatement detection as it measures agreement between continuous scores while penalising systematic over- or underestimation of claim strength; values closer to 1 indicate strong agreement, while lower values reflect miscalibration or inconsistent scoring. We also report mean absolute error (MAE) to quantify the magnitude of scoring deviations and Spearman's ρ to capture relative ranking consistency independent of calibration, providing complementary views of both calibration and ordering performance.

Model	CCC ↑	MAE ↓	ρ ↑
Text-only models			
Deepseek-V3.2	0.356	0.195	0.392
Deepseek-R1	0.463	0.201	0.544
GLM-4.6	0.385	0.240	0.490
Vision–language models (VLMs)			
InternVL3.5-8B	0.106	0.326	0.158
InternVL3.5-38B	0.347	0.161	0.360
InternVL3.5-30B-A3B	0.133	0.295	0.257
Qwen3-VL-8B	0.323	0.237	0.428
Qwen3-VL-32B	0.456	0.187	0.532
GPT-5-mini (low)	0.478	0.209	0.571
GPT-5-mini (high)	**0.493**	0.204	**0.587**
LLaVA-OV-1.5-8B	0.088	0.241	0.116
Ovis2-34B	**0.493**	**0.154**	0.509
GLM-4.5V	0.358	0.169	0.447

Table 3: Base model performance on overstatement detection utilising the claim-evidence sets.Higher is better for CCC and ρ, lower is better for MAE; **Bold** = best; underline = 2nd best.

5 Experimental results

Can Models Retrieve Intra-Paper Evidence? In the zero-shot setting, rerankers with generative or hybrid relevance modelling consistently outperform encoder-only models, suggesting that matching scientific claims to internal evidence benefits from richer semantic reasoning beyond embedding similarity. Among zero-shot methods, the E2Rank family performs strongly across metrics, with E2Rank-0.6B achieving competitive MAP and recall despite its smaller size, indicating that hybrid embedding–reranking approaches are effective for intra-paper evidence selection (as shown in Table 2).

Fine-tuning leads to substantial and consistent improvements across all evaluated models, confirming that evidence retrieval within scientific papers is a learnable task under the proposed automatic supervision. In particular, Qwen3-Reranker-8B exhibits the largest gains across MAP, Recall@K, and NDCG@K, suggesting that conventional reranking architectures are able to exploit the task-specific

Model	Setting	CCC ↑	MAE↓	ρ ↑
Qwen3-VL-8B	Base	0.323	0.237	0.428
	Text-only	0.529	0.156	0.593
	+Image	**0.578**	**0.153**	**0.649**
InternVL3.5-8B	Base	0.106	0.326	0.158
	Text-only	0.418	0.208	0.528
	+Image	0.479	0.191	0.557
LLaVA-OV-1.5-8B	Base	0.088	0.241	0.116
	Text-only	0.230	0.229	0.326
	+Image	0.317	0.205	0.392

Table 4: Performance before and after fine-tuning under three settings: Base (zero-shot), Text-only (fine-tuned using text inputs only), and +Image (fine-tuned with visual and text inputs). Higher is better for CCC and ρ, lower is better for MAE; **Bold** = best; underline = 2nd best.

supervision.

Can Models Detect Overstatements? Table 3 shows that strong text-only models can achieve competitive performance on overstatement detection, with DeepSeek-R1 substantially outperforming other text-only baselines and performing on par with several VLMs. This indicates that linguistic cues alone capture part of the signal, particularly for claims whose evidential support is primarily textual. VLMs such as GPT-5-mini (high) and Ovis2-34B achieve the strongest overall performance, exhibiting both higher agreement (CCC) and more reliable ranking (Spearman's ρ), while also maintaining lower MAE. Although absolute CCC values are modest, scores around 0.5 are expected for graded, subjective judgments and correspond to strong agreement given the metric's sensitivity to calibration. The variance among VLMs further suggests that current gains from visual inputs may be constrained by limitations in multimodal reasoning.

Table 4 analyses the effect of fine-tuning and visual grounding within the same model families. Fine-tuning substantially improves agreement and ranking consistency (CCC and Spearman's ρ), confirming that overstatement detection is learnable under the proposed supervision. Incorporating visual inputs further increases CCC and ρ across models, indicating that figures and tables provide complementary signal for assessing relative overstatement severity. Improvements in absolute error (MAE) are more model-dependent, particularly for InternVL and LLaVA; this is expected since reference scores are soft labels obtained by averaging multiple annotations, where gains are more reliably reflected in agreement and ranking than in exact score matching. Overall, multimodal fine-tuning enables smaller models to match or exceed the performance of much larger zero-shot counterparts.

6 Case Study

To illustrate how our framework behaves in practice, we conduct a case study on "*On Isotropy, Contextualization and Learning Dynamics of Contrastive-based Sentence Representation Learning*" (Xiao et al., 2023), with permission from the authors. Starting from a claim in the paper, we manually construct additional variants, with varying degrees of overstatement, by modifying only the claim's wording while keeping the supporting evidence fixed. This controlled manipulation allows us to directly probe whether the framework is sensitive to rhetorical amplification rather than differences in evidence. We utilise our best fine-tuned Qwen3-8B model, whose performance is reported in Table 4, for our case study.

Table 5 shows a clear progression in overstatement scores. The overstated variant employs absolute and theoretical language (e.g., "consistent", "formally prove") and introduces effects not directly measured in the experiments, resulting in the highest score. The partially overstated claim blends valid observations with unsupported generalisations, while the well-stated variant adheres closely to the reported results, using cautious language (e.g., "observations suggest") that matches the strength of the evidence. Overall, this case study highlights the core distinction targeted by our task: overstatement is not about factual incorrectness, but about how far a claim's wording stretches beyond its evidential footing. This observation is further supported by an analysis of linguistic certainty across different degrees of overstatement, as shown in Appendix F. Although illustrative, this case study provides a concrete example of the framework's behaviour in practice and complements the broader quantitative results reported in the previous sections.

7 Conclusion

We present a framework for detecting scientific overstatement by assessing whether the strength of claims in abstracts and introductions is proportionate to the evidence provided in the paper. By aligning claims with multimodal evidence and incorporating peer-review context during annotation, we capture graded differences in evidential sup-

Claim: Our analyses indicate a **consistent inward movement** of embeddings accompanied by **reduced anisotropy**, which **clarifies under-represented regions**. **Score: 0.6** **Reasoning:** The claim uses strong language ("consistent," "clarifies") that isn't fully supported by the evidence. The evidence shows a trend of inward movement and reduced anisotropy, but doesn't demonstrate a complete or consistent effect. The claim about clarifying under-represented regions is not directly supported by the provided data, which focuses on geometric changes and isotropy rather than semantic coverage. The evidence also lacks details on the specific regions being clarified, making the claim an overstatement of the findings.	**Claim:** We **formally prove** that contrastive learning induces embeddings to ***converge at the origin***. **Score: 0.85** **Reasoning:** The claim uses the strong term "formally prove," which implies a rigorous mathematical proof, but the evidence consists of empirical observations, visualizations, and correlations (e.g., L2-norm trends, UMAP plots) without any formal proof. The evidence shows that embeddings tend to move toward the origin and become more isotropic, but this does not constitute a formal proof of convergence. The claim overstates the strength of the evidence by implying a level of theoretical certainty that is not supported by the data.
Claim: We find that the embedding space is **directed towards the origin** during training, with **more areas now better defined**. **Score: 0.2** **Reasoning:** The claim accurately states that embeddings move toward the origin and become more defined, supported by evidence showing reduced anisotropy and better-defined areas.However, it slightly overstates by not mentioning the role of temperature and the specific conditions under which these changes occur, leading to minor exaggeration.	**Claim:** Our **observations suggest** that the movement of embeddings toward the origin is **one of several dynamics** contributing to the organization of the semantic space. **Score: 0.1** **Reasoning:** The claim is well-supported by the evidence, which demonstrates that embedding movement toward the origin is one of several dynamics contributing to semantic space organization. The evidence includes experiments showing the effects of temperature and pooling methods on isotropy, as well as visualizations of embedding changes. The claim does not overstate the evidence, as it accurately reflects the findings without exaggeration.

Table 5: A case study demonstrating different degrees of overstatement. Colour-coding highlights the key phrases in each claim that influence the model's judgement, and the full supporting evidence is provided in Appendix H.

port rather than binary factual correctness. Our experiments show that overstatement detection is learnable from these annotations and benefits from multimodal grounding, while also highlighting current limitations in model reasoning over visual evidence. Overall, this work reframes scientific rigour as a question of evidential proportionality and provides a foundation for tools that support clearer and more faithful scientific communication.

Limitations

Overall, the model produces explanations that align with the authors' own assessment. While our work is intended to encourage clearer, well grounded scientific communication, it could be misused. It is not intended to replace peer reviewing but to assist in the process. A high overstatement score should therefore be treated as a prompt for closer human review, not a basis for rejection.

Our approach is designed around the structure of scientific papers available in OpenReview, and naturally reflects the conventions and formatting typically found in this setting. As a result, the system may require adaptation when applied to other venues or scientific fields with different writing styles or evidence formats. In addition, our evaluations focus on alignment between claims and the evidence presented within a paper, rather than on broader scientific correctness, so the tool should be viewed as supporting clarity and grounding rather than making judgments about the overall quality of the work.

The current study focuses on NLP and ML venues, where claims are expected to be tightly grounded in empirical results. While the framework is discipline-agnostic in principle, as the same alignment-based criterion applies as long as claims and supporting evidence can be extracted, stylistic norms vary across fields and broader cross-disciplinary applicability remains an avenue for future work.

Acknowledgements

Joseph James is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

References

AAAI. 2025. AAAI Launches AI-Powered Peer Review Assessment System. https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/. Accessed: 27 November 2025.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *Preprint*, arXiv:2509.23661.

ARR. 2025. ACL Rolling Review - Responsible NLP Research Checklist. https://aclrollingreview.org/responsibleNLPresearch/. Accessed: 27 November 2025.

Sandeep B Bavdekar. 2015. Writing introduction: Laying the foundations of a research paper. *Journal of the Association of Physicians of India*, 63(7):44–6.

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2023. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*.

Prabhat Kumar Bharti, Tirthankar Ghosal, Mayank Agarwal, and Asif Ekbal. 2024. Peerrec: An ai-based approach to automatically generate recommendations and predict decisions in peer review. *International Journal on Digital Libraries*, 25(1):55–72.

Chu Sern Joel Chan, Aakanksha Naik, Matthew Akamatsu, Hanna Bekele, Erin Bransom, Ian Campbell, and Jenna Sparks. 2024. Overview of the context24 shared task on contextualizing scientific claims. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 12–21.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers.

Viviana Cortes. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes*, 23(4):397–423.

Mark De Rond and Alan N Miller. 2005. Publish or perish: Bane or boon of academic life? *Journal of management inquiry*, 14(4):321–329.

DeepSeek-AI. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI. 2025b. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, et al. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151*.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*, pages 101–109.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Team GLM-V, Wenyi Hong, Wenmeng Yu, and et al. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. 2025. Peer reviews of peer reviews:

A randomized controlled trial and other experiments. *PloS one*, 20(4):e0320444.

Abel Gustafson and Ronald E Rice. 2020. A review of the effects of uncertainty in public science communication. *Public Understanding of Science*, 29(6):614–633.

Ken Hyland and Feng Kevin Jiang. 2021. 'our striking results demonstrate…': Persuasion and the growth of academic hype. *Journal of Pragmatics*, 182:189–202.

ICLR. 2025. Assisting ICLR 2025 reviewers with feedback. https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/. Accessed: 27 November 2025.

ICML and ICLR. 2019. ICML/ICLR Workshop on Reproducibility in Machine Learning. https://sites.google.com/view/icml-reproducibility-workshop/home. Accessed: 1 January 2026.

Atiqa Iftikhar, Muhammad Khalil, and Iqra Asghar. 2025. Hedging and emphasis: The use of discourse markers in undergraduate research writing. *Indus Journal of Social Sciences*, 3(1):683–692.

Kristen Intemann. 2022. Understanding the problem of "hype": Exaggeration, values, and trust in science. *Canadian Journal of Philosophy*, 52(3):279–294.

Joseph James, Chenghao Xiao, Yucheng Li, and Chenghua Lin. 2024. On the rigour of scientific writing: Criteria, analysis, and insights. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6523–6538, Miami, Florida, USA. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Istvan Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr'as Gyorgy, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Boxi Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, Cj Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, J. Michael Wieting, Jonathan Lai, Jordi Orbay, Joe Fernandez, Joshua Newlan, Junsong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Ping mei Xu, Piotr Stańczyk, Pouya Dehghani Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Ardeshir Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vladimir Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab S. Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and L'eonard Hussenot. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Wei-Yu Kao and An-Zi Yen. 2024. How we refute claims: Automatic fact-checking through flaw identification and explanation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 758–761.

Tomoyuki Kawase. 2018. Rhetorical structure of the introductions of applied linguistics phd theses. *Journal of English for Academic Purposes*, 31:18–27.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Qi Liu, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, and Jiaxin Mao. 2025. E2rank: Your text embedding can also be an effective and efficient listwise reranker. *arXiv preprint arXiv:2510.22733*.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. 2025. AAAR-1.0: Assessing AI's potential to assist research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 40361–40383. PMLR.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

ML Reproducibility Challenge. 2025. The Machine Learning Reproducibility Challenge. https://reproml.org/. Accessed: 1 January 2026.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. 2025. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*.

Niklas Muennighoff, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

NeurIPS. 2025. Results of the NeurIPS 2024 Experiment on the Usefulness of LLMs as an Author Checklist Assistant for Scientific Papers. https://blog.neurips.cc/2024/12/10/results-of-the-neurips-2024-experiment-on-the-usefulness-of-llms-as-an-author-checklist-assistant-for-scientific-papers/. Accessed: 27 November 2025.

Made Nindyatama Nityasya, Haryo Wibowo, Alham Fikri Aji, Genta Winata, Radityo Eko Prasojo, Phil Blunsom, and Adhiguna Kuncoro. 2023. On "scientific debt" in NLP: A case for more rigour in language model pre-training research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8554–8572, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2025. *GPT-5 System Card*. Version dated August 13, 2025.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qiyao Peng, Chen Wang, Yinghui Wang, Hongtao Liu, Xuan Guo, and Wenjun Wang. 2025. Frontier-revrec: A large-scale dataset for reviewer recommendation. *arXiv preprint arXiv:2510.16597*.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20.

Team Qwen. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shruthan Radhakrishna, Aman Tiwari, Aanjaneya Shukla, Masoud Hashemi, Rishabh Maheshwary, Shiva Krishna Reddy Malay, Jash Mehta, Pulkit Pattnaik, Saloni Mittal, Khalil Slimi, Kelechi Ogueji, Akintunde Oladipo, Soham Parikh, Oluwanifemi Bamgbose, Toby Liang, Ahmed Masry, Khyati Mahajan, Sai Rajeswar Mudumba, Vikas Yadav, Sathwik Tejaswi Madhusudhan, Torsten Scholak, Sagar Davasam, Srinivas Sunkara, and Nicholas Chapados. 2025. Apriel-1.5-15b-thinker. *Preprint*, arXiv:2510.01141.

Wullianallur Raghupathi, Viju Raghupathi, and Jie Ren. 2022. Reproducibility in computing research: An empirical study. *IEEE Access*, 10:29207–29223.

Mizanur Rahman, Saadiyah Darus, and Zaini Amir. 2017. Rhetorical structure of introduction in applied linguistics research articles. *Educare*, 9(2).

Seema Rawat and Sanjay Meena. 2014. Publish or perish: Where are we heading? *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 19(2):87.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Françoise Salager-Meyer. 1994. Hedges and textual communicative function in medical english written discourse. *English for specific purposes*, 13(2):149–170.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

Marta Serra-Garcia and Uri Gneezy. 2021. Nonreplicable publications are cited more than replicable ones. *Science advances*, 7(21):eabd1705.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand. Association for Computational Linguistics.

Olga Stavrova, Bennett Kleinberg, Anthony M Evans, and Milena Ivanović. 2025. Scientific publications that use promotional language in the abstract receive more citations and public attention. *Communications Psychology*, 3(1):118.

ByteDance Seed Team. 2025. Seed-oss open-source models. `https://github.com/ByteDance-Seed/seed-oss`.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. 2025. Kimi-VL technical report. *Preprint*, arXiv:2504.07491.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. Evaluating large language models as expert annotators. In *Second Conference on Language Modeling*.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

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

Gang Wang, Qi Peng, Yanfeng Zhang, and Mingyang Zhang. 2023. What have we learned from openreview? *World Wide Web*, 26(2):683–708.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Melissa A Wheeler, Ekaterina Vylomova, Melanie J McGrath, and Nick Haslam. 2021. More confident, less formal: Stylistic changes in academic psychology writing from 1970 to 2016. *Scientometrics*, 126(12):9603–9612.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. 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 7053–7065, Online. Association for Computational Linguistics.

Chenghao Xiao, Yang Long, and Noura Al Moubayed. 2023. On isotropy, contextualization and learning dynamics of contrastive-based sentence representation learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12266–12283, Toronto, Canada. Association for Computational Linguistics.

Mingxin Yao, Ying Wei, and Huiyu Wang. 2023. Promoting research by reducing uncertainty in academic writing: a large-scale diachronic case study on hedging in science research articles across 25 years. *Scientometrics*, 128(8):4541–4558.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2025. Minicpm-v: A gpt-4v level mllm on your phone. *Nat Commun 16, 5509 (2025)*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *Preprint*, arXiv:2509.18154.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, Gelareh Mohammadi, and Aaron Quigley. 2025. A case study of scalable content annotation using multi-llm consensus and human review. *arXiv preprint arXiv:2503.17620*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

A LLM annotation

Model selection All models were obtained from Hugging Face. Specifically, we used ByteDance-Seed/Seed-OSS-36B-Instruct, openai/gpt-oss-120b, google/gemma-3-27b-it, ServiceNow-AI/Apriel-1.5-15b-Thinker, deepseek-ai/DeepSeek-R1-Distill-Qwen-32B, moonshotai/Kimi-VL-A3B-Instruct, openbmb/MiniCPM-V-4_5, and Qwen/Qwen3-VL-30B-A3B-Instruct. These models span diverse architectures and training paradigms, covering both text-only and vision–language reasoning models.

Tables 6 and 7 report analyses assessing the robustness of the annotation and aggregation procedure under leave-one-model-out settings, while Table 8 summarises the effect of incorporating peer-review context across different initial overstatement levels.

Excluded Model	Own	Text	Image
Seed-OSS-36B	0.9837	0.8421	—
GPT-OSS-120B	0.9801	0.7746	—
DeepSeek-R1-32B	0.9801	0.8368	—
Qwen3-VL-30B-A3B	0.9806	0.7536	0.7531
Gemma-3-27B-it	0.9822	0.8968	0.7553
Kimi-VL-A3B	0.9815	0.8597	0.9268
Apriel-1.5-15B	0.9892	0.9950	0.8497
MiniCPM-V-4.5	0.9823	0.9945	0.7526

Table 6: Krippendorff's α agreement between full and excluded-model consensus across annotation settings (all $p < 0.01$). **Own:** Authors original statement annotation. **Text:** Relevant text annotation. **Image:** Relevant visual annotation.

Model	Δ Mean	MAD	Welch p
Seed-OSS-36B	+0.0160	0.0305	< 0.01
GPT-OSS-120B	+0.0085	0.0239	< 0.01
Gemma-3-27B-it	−0.0191	0.0203	< 0.01
Apriel-1.5-15B	+0.0024	0.0182	0.3167
DeepSeek-R1-32B	+0.0105	0.0244	< 0.01
MiniCPM-V-4.5	−0.0110	0.0306	< 0.01
Qwen3-VL-30B-A3B	−0.0074	0.0280	< 0.01
Kimi-VL-A3B	−0.0035	0.0240	0.084

Table 7: Change in mean overstatement scores when excluding each annotator model. Δ Mean measures how much the overall average score shifts relative to the full-model baseline, while MAD (Mean Absolute Deviation) reflects the average absolute per-claim change in the aggregated scores.

Initial score band	$\Delta\mu$	$\tilde{\Delta}$	$\lvert\Delta\rvert$	$\uparrow$ / $\downarrow$ / $=$ (%)
Low (0.0–0.3)	+0.0978	+0.0625	0.1157	73.0 / 16.2 / 10.8
Low–Mid (0.3–0.5)	+0.0471	+0.0333	0.1107	58.1 / 34.9 / 7.0
Mid (0.5–0.7)	-0.0457	+0.0000	0.0696	22.3 / 41.1 / 36.7
High (0.7–1.0)	-0.1362	-0.0889	0.1475	11.7 / 79.3 / 9.0

Table 8: Impact of peer-review context stratified by initial (paper-only) overstatement score. Positive values indicate increased criticality after incorporating reviews, while negative values indicate reduced scores.

B Evidence retrieval: Training details

Full evidence retrieval dataset breakdown shown in Table 9. All models were tested and trained on a single A100 GPU, with hyperparameters provided in Table 11 Prompt used for LLM-based models is shown in Table 10. Full training details provided in Table 11.

Evidence Type	Train	Dev	Test	Total
Supporting	183,518	77,333	24,548	285,399
TEXT	146,243	62,724	20,693	229,660
IMAGE	37,275	14,609	3,855	55,739
Not-supporting	246,001	113,081	37,490	396,572
TEXT	195,942	91,955	31,815	319,712
IMAGE	50,059	21,126	5,675	76,860

Table 9: Detailed breakdown of evidence.

```
You are an evidence verification assistant.
Given a claim and a document, determine if the
document provides supporting evidence for the
claim.
INSTRUCTION: Does the following document
provide supporting evidence for the claim?
```

Table 10: Prompt for finetuning LLM based reranker models.

Model selection All models were obtained from Hugging Face. Specifically, we used ALIBABA-NLP/E2RANK-(0.6B,4B,8B), QWEN/QWEN3-RERANKER-(0.6B,4B,8B), SENTENCE-TRANSFORMERS/ALL-MINILM-L6-V2, BAAI/BGE-RERANKER-V2-M3, ALIBABA-NLP/GTE-MULTILINGUAL-RERANKER-BASE, and GRITLM/GRITLM-7B.

C Overstatement detection: Training details

All models were tested and trained on a single A100 GPU, with hyperparameters provided in Table 13. Prompt used for fine-tuning Task 2 is shown in Table 12. Full training details provided in Table 13.

Hyperparameter	
epochs	3
batch_size	16
gradient_accumulation_steps	4
gradient_checkpointing	True
optim	adamw
learning_rate	3e-5
weight_decay	0.01
max_grad_norm	1.0
warmup_ratio	0.1
logging_steps	500
eval_steps	1000
eval_strategy	steps
save_steps	1000
EarlyStoppingCallback	5

Table 11: Training hyperparameters for evidence retrieval.

```
You are a model specialized in assessing
overstated claims using text and image evidence.
You must score each claim based on how
overstated or exaggerated it is with respect
to the evidence,on a continuous scale from 0
to 1 where 0 means well-stated and 1 means
overstated.
Provide the final score as <score>value</score>
followed by a brief reasoning.
```

Table 12: Prompt for finetuning for overstatement detection.

Hyperparameter	
epochs	3
batch_size	1
gradient_accumulation_steps	16
gradient_checkpointing	True
optim	adamw
learning_rate	3e-5
weight_decay	0.01
max_grad_norm	1.0
warmup_ratio	0.1
logging_steps	500
eval_steps	1000
eval_strategy	steps
save_steps	1000
EarlyStoppingCallback	5

Table 13: Training hyperparameters for overstatement detection.

Model selection For larger open-source and closed-source models (ZAI-ORG/GLM-4.5V, ZAI-

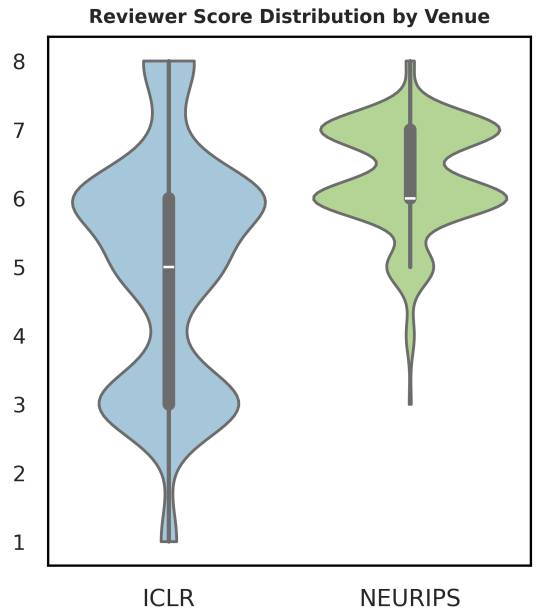


Figure 1: Reviewer rating distribution for ICLR and NeurIPS.

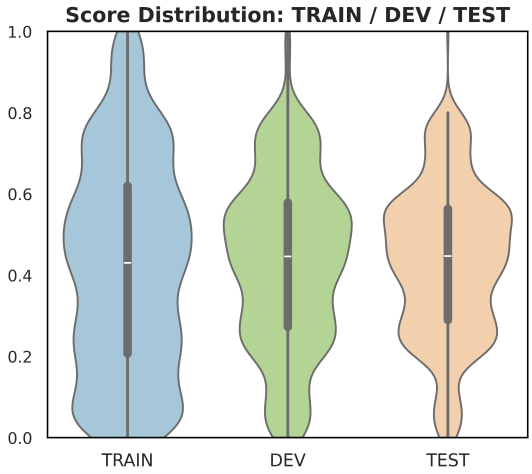


Figure 2: Score distributions for each split

ORG/GLM-4.6, DEEPSEEK-V3.2. and GPT-5-MINI-2025-08-07), we utilised APIs to obtain model outputs through their hosted inference endpoints, as local deployment was not feasible due to GPU memory limitations. For all other models we utilised a single A100 for the following models; QWEN/QWEN3-VL-32B-INSTRUCT, OPENGVLAB/INTERNVL3_5-38B-HF, OPENGVLAB/INTERNVL3_5-30B-A3B-HF, and AIDC-AI/OVIS2-34B. For finetuning we utilised LMMS-LAB/LLAVA-ONEVISION-1.5-8B-INSTRUCT, OPENGVLAB/INTERNVL3_5-4B-HF, and QWEN/QWEN3-VL-8B-INSTRUCT.

D Human evaluation

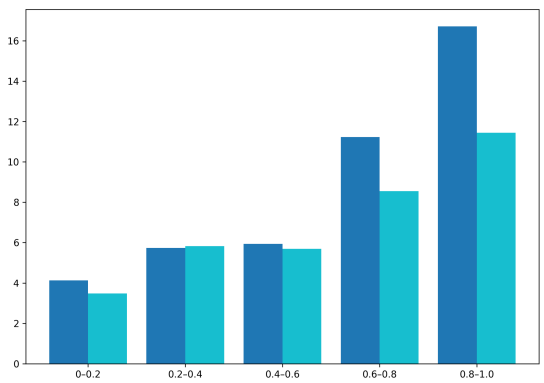
Human evaluators were provided 20GBP in Amazon gift cards per hour to complete the evaluation, which is above the minimum wage in the UK. Further allowing for up to 3 hours of work to allow time for a thorough analysis of the claim and evidence.

E Dataset details

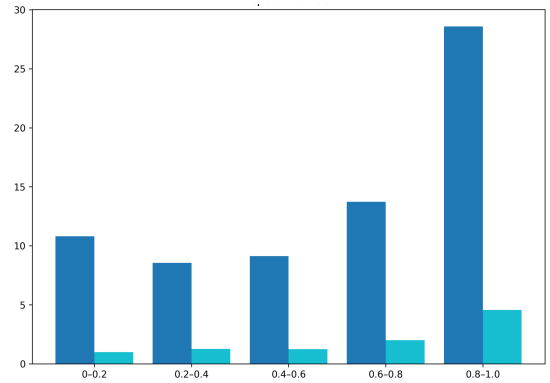
F Certainty Measure

We investigate how varying degrees of overstatement influences the certainty of the claim (Pei and Jurgens, 2021). Figure 3 shows that claims use mainly confident language with uncertainty aspects being uncommon, reflecting the standard that researchers present their findings with confidence and clarity (Salager-Meyer, 1994; Cortes, 2004).

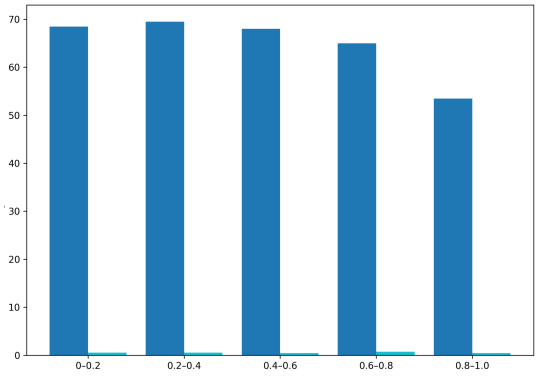
The observed patterns indicate that increased overstatement is driven by a greater use of probability certainty (e.g., *will*, *guaranteed*), extent certainty (e.g., *fully*, *completely*, *exactly*), and number certainty (e.g., precise or absolute quantification), each of which serves to present claims as more definitive than is typically warranted. In contrast, framing certainty (e.g., *show*, *demonstrate*, *verify*) exhibits a slight decline, suggesting a reduced reliance on evidential framing in favour of more assertive language. Suggestion and condition aspects occur in fewer than 5% of claims and are therefore excluded from further analysis. These findings align with recent work on undergraduate theses, which reports a growing reliance on intensifiers across academic levels to strengthen claims; when paired with insufficient hedging, this tendency has been shown to undermine perceived credibility (Iftikhar et al., 2025). More broadly, this pattern is consistent with evidence that hedging language and other uncertainty-marking devices have become less common in academic writing in recent years (Wheeler et al., 2021; Yao et al., 2023).



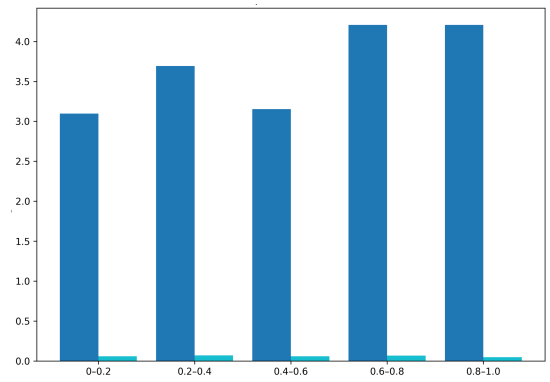
(a) Extent



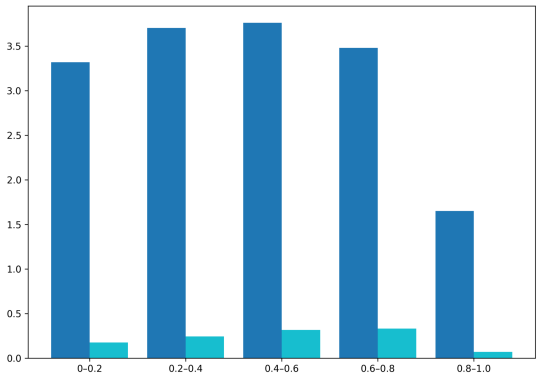
(b) Number



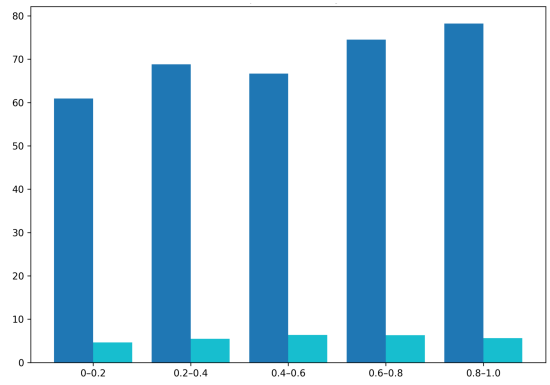
(c) Framing



(d) Condition



(e) Suggestion



(f) Probability

Figure 3: Distribution of certainty and uncertainty linguistic aspects across overstatement score bins. Each subfigure corresponds to a certainty category (Extent, Number, Framing, Condition, Suggestion, Probability). Dark blue bars indicate certainty expressions, while light blue bars indicate uncertainty expressions.

G Prompts for data processing

```
You will be provided with the abstract and introduction of an academic paper along with a specific
sentence from the paper.  Your task is to determine whether the given sentence represents an
original claim introduced by the authors that is directly relevant to the contribution or selling
points of the paper.

Labels:
original_statement: The sentence explicitly presents a novel claim, finding, or result that is
directly relevant to the key contributions of the paper. It reflects what the authors are aiming
to promote or highlight as a significant contribution.

not_original_statement: The sentence mainly provides background information, references prior
work, describes common knowledge, or includes general context not directly tied to the unique
contributions of the paper.

The abstract and introduction of the paper:
Abstract:
{ABSTRACT}

Introduction:
{INTRODUCTION}

The sentence you are about to annotate:
{SENTENCE}

You should:
1. Carefully review the context of the paper (abstract and introduction) and the given sentence.
Then briefly justify whether the sentence is an original_statement or not_original_statement (up
to 100 words).
2. Provide the final annotation label in the format: <Label>{your_label}</Label>
```

Table 14: Prompt for Own statement labelling.

```
You will be given a claim and a list of sentences. Your task is to identify the sentences that
support the claim.

A sentence supports the claim if it:

- Directly provides evidence (e.g., experimental results, analysis, conclusions).
- Builds upon the claim by providing relevant context (e.g., background information).

A supporting sentence must not:

- Be a duplicate or paraphrase of the claim.
- Be incomplete.
- Contain text that appears to be part of an OCR-extracted table or figure (e.g., columns of
numbers, symbols, "Table 1", or values not from a sentence). Such lines should always be ignored.

The sentences are numbered, and you should return only the numbers of the supporting sentences.

Claim:
{CLAIM}

Sentences to evaluate:
{NUMBERED SENTENCES}

Instructions:
Carefully review the claim and sentences. Provide a brief justification (≤100 words) for which
sentences support the claim.
If multiple sentences support the claim, list each number on a new line. If no sentences support
the claim, return an empty <Label> tag.
Provide the final annotation label in the format:
<Label>
{sentence numbers}
</Label>
```

Table 15: Prompt for text evidence extraction.

```
You will be provided with a research claim and a {FIG_TYPE} (figure or table) extracted from an
academic paper.
Your task is to determine whether the visual content is relevant to the claim — that is, whether
it provides evidence or context supporting the claim.

A visual is relevant if it:

- Directly provides evidence (e.g., experimental results, analysis, conclusions).
- Builds upon the claim by providing relevant context (e.g., background information).

A visual is not relevant if it:

- Contains no data or analysis tied to the claim.
- Shows unrelated or generic material.
- Is incomplete, unreadable, or too vague to judge its relevance.

Labels:
relevant: The visual supports or builds upon the claim.
not_relevant: The visual is unrelated to the claim.

Claim:
{CLAIM}

Visual information:
Type: {FIG_TYPE}
Caption: {CAPTION}
Visible text: {IMAGE_TEXT}

Instructions:
1. Carefully review the claim and the visual.
2. Briefly justify (≤100 words) whether the visual is relevant or not.
3. Provide the final label in this format:
<Label>{relevant OR not_relevant}</Label>
```

Table 16: Prompt for image-based evidence extraction, used for figures and tables.

Your role is to assess the degree to which a claim is overstated based on the available evidence.

"Overclaiming" refers to rhetorical exaggeration: when the wording or framing of a claim amplifies its strength beyond what the paper's own evidence supports.
It concerns rhetorical and linguistic inflation rather than factual correctness.

The Input Information will include:
1. Original Claim: The claim under evaluation.
2. Evidence: Research findings, including figures, tables, or other relevant data supporting the claim.

Optional. Review comment: Reviewer feedback relevant to the claim's validity.

Evaluate the claim against the provided evidence. Assign a score from 0 to 1 representing the degree of exaggeration using the following scale:
0.0: The claim contains no exaggeration and fully aligns with the evidence.
Values closer to 0: Minor exaggeration or slight over-interpretation.
Values closer to 1: Substantial exaggeration beyond what the evidence supports.
1.0: Major exaggeration or strong misrepresentation of the evidence.

Justification: Provide a concise explanation that includes:
Instances of exaggerated wording, insufficient experiments, lack of experimental details, gaps in knowledge, weak grounding in evidence, or missing limitations.
Direct references to the relevant evidence supporting your reasoning.
If a review comment is included, consider relevant points but do not mention or reference the review.
Do not mention or restate the score in the justification.

The claim to be assessed is:
{CLAIM}

The review comment to be evaluated is:
{REVIEW}

The evidence to be evaluated is:
{EVIDENCE}

You should:
1. Review the claim and the text and image evidence. Summarize how the evidence influences your evaluation of the claim and briefly explain whether the claim is well-stated or overstated on the 0–1 scale (up to 100 words).
2. Provide the final score in the format: <score>{score}</score>.
3. Provide your justification in the format: <justification>{justification}</justification>.

Table 17: Prompt for Overstatement label annotation.

H Case study evidence

Evidence (Input)

Geometrically, the embeddings of tokens are pushed toward the origin in the output layer of a model, compressing the dense regions in the semantic space toward the origin, making the embedding space more defined with concrete examples of words (see also Figure 1), instead of leaving many poorly-defined areas (Li et al., 2020). We provide a visualization of embedding geometry change in Figure 1. We suggest that this range plays a main role in making the entire semantic space isotropic. We find that temperature affects making embeddings isotropic: to push in-batch negatives to the lower bound, the temperature needs to be twice as large than to push them to the upper bound. Connecting this to our finding on high intra-sentence similarity, we observe that given a sentence/document-level input, certain semantic tokens drive the embeddings of all tokens to converge to a position, while functional tokens follow wherever they travel in the semantic space. Their performance on L2-norm is also well-aligned, again showing strong correlation between isotropy and L2-norm in the training process utilizing contrastive loss. Further, with limited space in the now compressed space, inputs have now learned to converge to one another to squeeze to a point while keeping its semantic relationship to other examples. We perform UMAP dimensionality reduction on embeddings provided by models up to 1000 step to preserve better global structure, and visualize only vanilla and 200-step embeddings. Specifically, for anisotropy baseline, temperature being too low even augments the vanilla model's unideal behavior, and the same applies for L2-norm, by that temperature being too low actually pushes the embeddings even further from the origin. By contrast, mean pooling and max pooling demonstrate a faster convergence, with mean pooling being most promising on isotropy. For instance, removing the top 1 dominant dimension of minilmfinetuned seems to not affect the embeddings' relative similarity to one another at all, preserving an r^2 of .998. Figure 1: Expanded semantic space produced by contrastive learning (CL), visualized with UMAP. At the beginning of training, all embeddings occupied a narrow cone. After 200 steps of fine-tuning with a contrastive loss, they spread out to define a larger semantic space. Firstly, we present the centered property we are measuring, anisotropy. We showed the theoretical promise of uniformity brought by contrastive learning through measuring anisotropy, complemented by showing the flattened domination of top dimensions. Higher self-similarity indicates less contextualization. The central question posed in this paper revolves around the mechanism involved in the contrastive learning process that diminishes anisotropy, leading to an isotropic model. Figure 7 further validates this through showing that higher temperatures compress the semantic space in general, pushing instances to the origin. Given a token x, we denote the set of token embeddings of x contextualized by different contexts in corpus S as SX. We find that with optimal hyperparameters, the representations go through less change after 200 steps. We first use the vanilla mpnet to encode the STSB subset we have constructed.

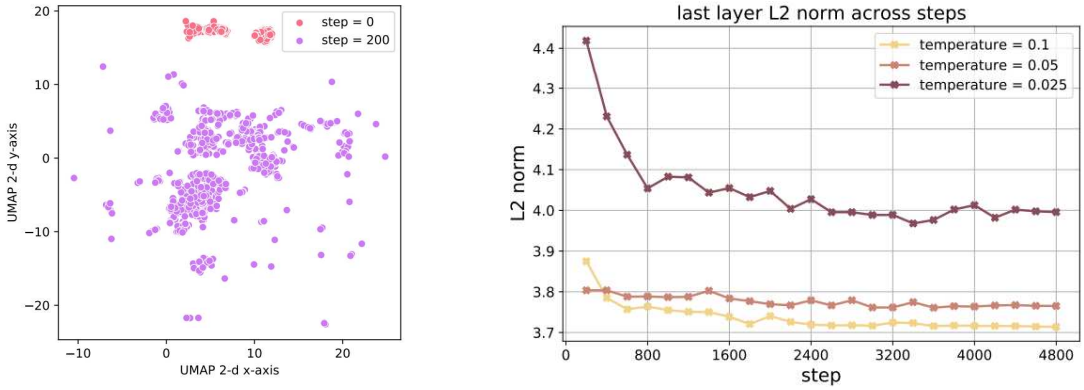


Table 18: Evidence utilised for case study retrieved using our fine-tuned Qwen reranker on the top 20 relevant sentences.