Universiteit Leiden
The Netherlands

Bachelor Computer Science & Datascience and Artificial Intelligence

Surface or Substance? A Controlled Experiment on How Contextual Signals Sway LLM Peer Review Judgements in CS

Hussein Al-Amodi

First supervisor:
Dr. T.D.P. Heyman
Second supervisor:
Evert van Nieuwenburg

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

20/02/2026

Abstract

Large language models (LLMs) are increasingly involved in peer review. However, the sensitivity of LLMs to text-level surface signals has been relatively untested. Prior studies have established various biases of LLMs associated with author identity, prompt framing, verbosity, and positioning during evaluation. However, the manipulation of an attribute of the text being reviewed in the absence of changes to its scientific content has not yet been explored. We test whether LLMs, specifically GPT-4o-mini, exhibit biases towards the presence/absence of quantitative signals in abstracts. Every abstract was assessed by GPT-4o-mini on three baseline evaluations and three post-manipulation evaluations, scoring each abstract on a scale of 1-100. Manipulation involved the removal of quantitative framing from abstracts that originally featured it or its introduction into abstracts without such framing. Removing quantitative language yielded a reduction in scores ($\Delta = -1.97$, $p = 0.002$, $d = -0.573$), while adding it produced an improvement in scores ($\Delta = +1.64$, $p = 0.034$, $d = +0.431$). The overall cross-condition difference of -3.61 points ($p < 0.001$) is consistent with the bidirectional effect. Altogether, the findings support the hypothesis that GPT-4o-mini has learned a heuristic connecting quantitative content signals to text quality based on its training data. As applied to conference screening systems based on LLMs, such biasing will confer an intrinsic advantage to papers featuring quantitative specificity regardless of their scientific contribution. Importantly, this biasing signal is inherent to the text itself and thus cannot be mitigated by author anonymization. Code is available here[1].

[1] https://github.com/Huss-Alamodi/surface-or-substance

Contents

1 Introduction

Top conferences in computer science and AI now receive tens of thousands of submissions each year. The number of qualified reviewers has not kept pace (Kim et al., 2025). NeurIPS and ICLR received roughly 18,000 submissions in 2024, and ICML received approximately 23,000 submissions in 2026[2]. The reviewer pool cannot handle this volume. Conferences compel authors to serve as reviewers, which sometimes means reviewers evaluate papers outside their direct expertise. Review quality varies widely as a result.

These conditions have made AI-assisted peer review an active area of consideration, and LLMs are now being deployed in conference workflows in several distinct roles. The ICML 2026 LLM policy now requires authors to disclose any AI assistance used during the review process[3]. The policy treats the involvement of LLMs as a structural shift in how peer review operates, with other venues likely to follow. It also regulates how AI may be used in the review process itself, including the disclosure of any AI assistance used by reviewers.

Liang et al. (2024) found that between 6.5% and 16.9% of sentences in reviews at ICLR 2024, NeurIPS 2023, CoRL 2023, and EMNLP 2023 had been substantially modified by LLMs. These changes were associated with lower reviewer engagement and greater homogeneity between reviews. Latona et al. (2024) found that at least 15.8% of ICLR 2024 submissions had been reviewed by AI. Submissions near the acceptance threshold that received an AI-assisted review were accepted 4.9 percentage points more frequently than similar submissions without AI assistance. Importantly, these facts derive from observational data gathered during actual review procedures, showing that LLMs indeed influence peer reviews and making the study of LLM evaluations of papers relevant.

Currently, there are at least four types of bias that LLMs exhibit when evaluating papers. The first type involves information presentation bias in the evaluated text: Saito et al. (2023) found that GPT-4 prefers longer inputs even when quality is equal. A separate type of bias arises from the evaluation prompt rather than the evaluated text: Hwang et al. (2026) showed that LLM evaluators shift their verdicts when the prompt uses a positive rather than a negative predicate. The distinction matters because prompt-side biases can in principle be mitigated by changing the evaluation template, while input-side biases cannot. A third type concerns the evaluation mechanism itself. LLM judges can experience bias where their preferences in pairwise or list-wise comparisons may shift depending on where a candidate appears in the prompt rather than on content quality alone. Zheng et al. (2023) identified this position bias in LLM-as-a-judge settings, and Shi et al. (2025) showed that these positional effects are systematic within evaluators, but the magnitude of the effect varies across evaluators and tasks. A fourth type concerns author identity. Three of four tested LLMs assigned higher ratings to papers attributed to male authors and authors from prestigious universities compared to the same papers presented anonymously, with effect sizes exceeding the gap between top-25 and 51-75 ranked journals (Pataranutaporn et al., 2025).

This raises the question: what properties of input text could trigger such biases in the context of peer review? Papers vary on many dimensions at once, so observational studies cannot isolate the mechanism due to the numerous variables associated with each paper (Latona et al., 2024; Liang et al., 2024). Demographic studies vary external demographic factors, not text (Pataranutaporn

[2] https://www.ucl.ac.uk/engineering/news/papers-accepted-icml-2026

[3] See https://icml.cc/Conferences/2026/LLM-Policy

et al., 2025). Prompt studies test instruction sensitivity, not content sensitivity (Hwang et al., 2026). Length studies cannot separate verbosity from quality (Saito et al., 2023). None of these varied individual features of the evaluated text while holding everything else constant. In this project, we manipulate a specific textual property while leaving the underlying scientific semantics unchanged. In brief, we test whether the presence or absence of quantitative framing in the abstract, the numerical expressions through which the paper's results are reported, affects the LLM-assigned quality score.

We define quantification along two components. Quantitative specificity refers to exact numbers, percentages, and effect sizes such as a reported 23% improvement with $p < 0.01$. Explicit scale claims refer to dataset size, model parameters, and compute requirements, such as a statement that a model was trained on 1.4M images with 175B parameters. Both components are common in CS/AI abstracts because the field expresses results numerically. Accuracy scores, benchmark comparisons, dataset sizes, and parameter counts form the core evidence of a paper's contribution. A reviewer evaluating an abstract relies on these numerical signals to assess the reported work. We hypothesise that removing these components will decrease LLM-assigned scores and that adding plausible versions of both where both are absent will increase scores, even when the underlying scientific content remains unchanged.

2 Background and Related Work

2.1 The Peer Review Landscape

The scale of modern AI research has outgrown the available reviewer pool. Major AI conferences now receive tens of thousands of submissions annually, yet the number of qualified reviewers has not kept pace. Many researchers decline review invitations because they are overloaded with their own submissions, teaching, and institutional duties. The result is a systemic shortage that forces conferences to rely on an increasingly smaller and less diverse pool of reviewers (Kim et al., 2025; Lerback & Hanson, 2017).

Several major machine learning conferences such as ICML and NeurIPS have taken steps to counter this problem by making it mandatory for all papers submitted to such conferences to have one of their authors act as a reviewer in the same cycle. This obligation does not increase the total pool of qualified reviewers because the same researchers who submit papers are the ones who review them. Instead, it creates a circular dependency: the same authors are both producers and consumers of reviews. Whether this dependency encourages less thorough evaluations is a structural concern, but the obligation itself is best understood as a response to the reviewer shortage rather than as a mechanism that directly degrades review quality. The problem for many reviewers is that they are faced with the difficult task of having to evaluate more papers than they are able to thoroughly analyse. This inevitably means that the quality of the decision decreases, the process takes longer, and the system becomes prone to superficial signals. Given that said reviewer has little time to conduct a comprehensive analysis, the name of the author, the institution they are affiliated with, and the title of the paper become more important. Additionally, a reviewer who receives a rushed or biased review may, consciously or not, apply the same standard to subsequent papers they review.

A double-blind process aims to eliminate bias based on knowing who the authors are and their affiliations, thus preventing an advantage for famous researchers or well-known institutions. In principle, this method helps mitigate bias in the form of non-content signals, yet in reality, this method is partly effective as reviewers can easily determine who the author is and what institution they are affiliated with through publicly available preprints (Pataranutaporn et al., 2025). As mentioned by Liang et al. (2024), there are two measurable dimensions between reviews with LLMs-generated text and reviews without them: lower reviewer engagement and higher uniformity in the reviews' content. The above findings are consistent with the interpretation that the usage of LLMs somehow changes the way reviewers conduct their business. Nevertheless, there is a lack of evidence proving that LLMs have a causal impact on engagement. In addition, according to Latona et al. (2024), at least 15.8% of all reviews presented at ICLR 2024 were generated using LLMs, which led to a 4.9% increase in the chances of acceptance of marginal papers.

2.2 LLM-as-a-Judge: Framework and Reliability

(Zheng et al., 2023) introduced MT-Bench, where they showed that the use of GPT-4 as an evaluator achieves more than 80% accuracy compared to human evaluators. As a result, LLMs became a plausible substitute for human evaluators and have since been used to evaluate chatbots, assess quality of summaries, and perform safety evaluations. Unfortunately, while LLMs offer many

benefits due to their ability to process natural language and apply learned criteria, they are equally susceptible to systematic bias.

LLM evaluators favour outputs from their own generative distribution. (Panickssery et al., 2024) demonstrated that LLMs recognise text produced using other models belonging to the same family and give it higher ratings. The self-preference bias was mathematically formulated by (Wataoka et al., 2024), whereby the higher the quality score for the text, the less perplexity of the text with respect to the parameters of the same model. LLMs also demonstrate position bias (Shi et al., 2025), whereby LLMs award higher scores to evaluated responses based on prompt position or formatting (Zheng et al., 2023), regardless of its content.

How LLMs respond to the features of the text they judge creates another set of problems. (Saito et al., 2023) found that GPT-4 consistently prefers longer responses despite the absence of quality improvement in the additional length, with a verbosity bias metric of 0.328. This metric ranges from 0 to 1, where 0 means the model shows no preference for length and 1 means it always prefers the longer option regardless of quality. A score of 0.328 indicates a moderate but measurable tendency. In practical terms, if two papers of equal quality are evaluated in a pairwise comparison, the longer abstract has roughly a 1 in 3 chance of being preferred over the equal quality shorter abstract solely because of its length. The bias is unique to LLMs, as human alignment drops when humans prefer the shorter response, suggesting that LLMs apply a "longer is better" heuristic. (Hwang et al., 2026) examined framing bias among LLMs by evaluating the effect of using positive and negative predicates in prompts, a bias that operates on the evaluation template rather than the evaluated text itself. Across all 14 LLMs tested, verdicts became systematically biased in different framings with inconsistency rates (the rate at which the model flips its judgment when we reframe the input) varying between 5.7% and over 66%.

However, the evaluation of LLMs is not limited to the mechanics and stylistics of generated text but also includes attribution. (Pataranutaporn et al., 2025) systematically varied author characteristics across 330 economics papers while holding content constant. Three out of the four tested LLMs assigned significantly higher ratings to papers attributed to top male authors and elite institutions in comparison to the same papers presented anonymously. The size of the effect was considerable. In some cases, the boost from author affiliation exceeded the difference between papers in top-25 and 51-75 ranked journals. Whether this represents a distinct LLM-specific bias or an inherited one depends on how the model acquired the association. Given that training of the LLM is performed based on scientific literature, journal reviews, and editor choices, it encodes the connection that exists in reality as well. A model trained using this corpus will reflect the correlations present in those texts. Crucially, the mechanism of this inherited bias differs from the human institutional bias it superficially resembles. A human reviewer may consciously or unconsciously associate elite institutions with higher quality based on direct experience or on the institution's reputation, both of which are causally upstream of the paper's actual quality. An LLM has no such experience or reputation. It has only the statistical patterns in its training data, where papers from elite institutions receive more citations and higher review scores on average. Whether this constitutes a bias depends on the comparison that is being made: against a reviewer who sees nothing other than the content of the paper, the LLM looks biased; against the training data, the LLM looks true to its source.

The breadth of these findings has motivated systematic taxonomies. (Ye et al., 2024) identify and

quantify 12 distinct categories of bias across multiple LLM judges, including authority bias, gender bias, beauty bias, and misinformation oversight bias. They find statistically significant deviations from expected unbiased behaviour in every category, across multiple model architectures. This consistency across studies, models, and tasks suggests that LLM evaluator invalidity is a structural property of current LLM-based evaluation, not a collection of isolated edge cases. The deviations from the expected baseline should more precisely be described as a loss of construct validity: the LLM evaluator's scores correlate with features other than the construct of interest, even when controlling for content.

2.3 Quantitative Framing as a Signal of Quality

None of the cited studies modifies any of the surface-level properties of the evaluated text without modifying the scientific content of said text. To clarify, surface-level properties of a text refer to properties that affect how an abstract expresses the scientific claims without affecting the claims themselves. Our experiment manipulates the surface-level properties of a text, without manipulating the scientific claim itself. Our modifications to the text do not affect the methodology employed, the conclusions reached, and the logic behind the claims. They merely modify how those claims are numerically formulated. Quantitative framing is one example of surface-level properties of a text. The existing literature on LLM influence in peer review falls into two strands that both leave a gap our study addresses. The first strand studies LLM-as-evaluator biases, where the LLM is the entity that scores the work. This work includes verbosity bias (Saito et al., 2023), prompt framing bias (Hwang et al., 2026), position bias (Shi et al., 2025; Zheng et al., 2023), and author identity bias (Pataranutaporn et al., 2025). None of these studies vary the text of the evaluated abstract while holding its scientific content constant. The second strand studies LLM-assisted reviewing by humans, where the LLM helps a human reviewer write or edit their review. (Liang et al., 2024) and (Latona et al., 2024) fall into this category, using observational data on review outcomes. These studies cannot isolate the effect of any textual property of the evaluated paper because the text being evaluated and the review being written are not under experimental control. Our study sits at the intersection: it asks the LLM to act as the evaluator, and it manipulates a property of the evaluated text while leaving the scientific content unchanged.

Our analysis investigates quantitative framing as a surface-level property. As meta-science evidence suggests, humans use the presence of quantification as a marker of scientific rigour. Indeed, as was shown by (Garcia-Costa et al., 2022), human peer reviewers explicitly call for more quantitative elements during peer reviews. In other words, the number of quantitative statements in articles appears to increase with progressing peer review stages. (Thelwall et al., 2023) found that articles that feature denser method sections written using quantitative language are scored higher during quality evaluations. Precise numbers are positively evaluated, while lack thereof is negatively penalized. The fact that this pattern operates in human review is what motivates the question of whether LLMs, trained on the same scientific literature, inherit the same heuristic.

In case this connection is embedded into scientific texts used for training LLMs, the evaluator would employ a similar heuristic. LLMs are trained on huge collections of scientific literature, in which the presence of quantitative content correlates with being accepted and getting good reviewer scores. According to the meta-study by (Li et al., 2024), such implicit associations can be embedded into

the evaluation function of LLM-as-judges algorithms. In other words, an LLM may rate an article lower if an edited version lacks its results, despite the fact that the actual contribution remains the same.

3 Research Question

No prior study has manipulated a surface-level textual property of a paper abstract while holding the underlying scientific content constant. The existing evidence is either observational (Liang et al., 2024) or varies metadata and prompts rather than the evaluated text itself ((Pataranutaporn et al., 2025); (Hwang et al., 2026)). To address this gap, we ask:

Does framing quantitative specificity and the inclusion of explicit scale claims affect LLM quality scores beyond the underlying scientific content?

A positive answer would establish that LLM evaluators encode text-level quality signals that are independent of scientific substance, with implications for any system that uses LLMs to triage or evaluate research.

4 Method

We conducted a controlled experiment to test whether quantitative framing in paper abstracts affects LLM-assigned quality scores when the scientific content remains unchanged. The experiment evaluates abstracts only, not full submissions. Each abstract, drawn from the ICLR 2024 review cycle, was evaluated by GPT-4o-mini both in its original form (baseline) and after a manipulation that altered the quantitative language. Two types of quantitative framing were manipulated: quantitative specificity (exact numbers, percentages, effect sizes) and explicit scale claims (dataset size, model parameters, compute requirements). The design follows a within-subject structure in which each abstract serves as its own control, and a difference-in-differences comparison across two conditions tests whether the effect is consistent across different paper populations.

4.1 Research Design

The design of the experiment includes two variables. The first one relates to whether the original abstract included both quantitative specificity and scale claims or not. Abstracts in the former case were placed into the QUANT-REMOVED condition, and abstracts in the latter case were placed into the QUANT-ADDED condition. The second variable relates to the evaluated form of the abstract; either the original text as retrieved, or a manipulated version in which quantitative language has been altered. Crossing these two variables provides us with a 2×2 within-subject design.

For the purpose of analysis, three quality bands have been established according to original paper ratings provided by human reviewers: rejected, if the average score was below 5; borderline, if the score ranged between 5 and 7; accepted, if the score was 7 and more. These bands are consistent with the reviewing practice adopted by ICLR in 2024. It is worth noting that these review scores are based on the entire paper, not just the abstract. The per-paper human quality estimate used throughout the analysis is the average of the original ICLR 2024 reviewer scores for that paper, where each paper received reviews from multiple ICLR reviewers.

In total, two populations of abstracts have been used. For QUANT-REMOVED condition, abstracts initially having both quantitative specificity and scale claims were selected. The quantitative language was removed while maintaining the scientific contribution made. For QUANT-ADDED condition, papers having neither of the characteristics mentioned were used. For this condition, we used Perplexity to generate plausible quantitative statements based on the full paper and manually verified them against the original text. We expect scores to decrease in the QUANT-REMOVED condition and increase in the QUANT-ADDED condition.

4.2 Classification and Sampling

The sample was drawn from roughly 7,000 abstracts submitted to ICLR 2024 review cycle through OpenReview. ICLR 2024 was chosen since its review data are available publicly in OpenReview, and the review cycle ended past October 2023, which is the reported cutoff date for GPT-4o-mini's training data. The focus on papers submitted post-training minimizes the possibility of the model's

exposure to the content of those papers during the training phase.

We used an LLM-based classifier to label each of the 7,000 papers by two binary attributes, (i) quantitative specificity and (ii) scale claims in its abstracts, in order to define whether a paper satisfies the conditions of the experiment. We ran the classification on the entire set of 7,000 papers. The classifier relied on GPT-4o-mini with a predefined output structure.

For each category (papers with quantitative specificity and scale claims vs. papers without those properties), we selected a random sample of 90 papers for manual review (30 papers in each of three quality bands). Pilot tests on 10 papers confirmed sufficient within-condition variance to detect differences of at least 2 points on a 1-100 scale.

In QUANT-REMOVED, all 90 randomly sampled papers were manually reviewed, and 36 met the experimental criteria upon human inspection. In QUANT-ADDED, manual review was constrained to 31 papers due to time limits, of which 27 met the criteria. The 27 surviving papers are the same set used in the QUANT-ADDED condition described in Section 4.3; the classifier's "no quantitative framing" output for these 31 papers was checked by human reviewers before inclusion. The acceptance rate is high in QUANT-ADDED ($27/31 = 87\%$) because the classifier is reliable at identifying papers that lack quantitative content, but human inspection was still required to confirm the absence of any quantitative framing that the classifier might have missed. The automated classifier had condition-specific error patterns. In QUANT-REMOVED, the classifier's "both" prediction was highly reliable (34 of 36 included papers were correctly flagged as having both attributes), but the "quant only" and "scale only" predictions had high false-positive rates (19 and 23 papers respectively were flagged as having one of the two attributes but rejected on manual review because they did not satisfy the joint criterion). In QUANT-ADDED, the false-positive rate was $4/31 = 13\%$ (4 of the 31 reviewed papers were classified as "no quantitative framing" but excluded on manual review). In both conditions, manual validation served as the ground-truth filter for sample inclusion. The final sample consisted of 36 papers in QUANT-REMOVED (16 rejected, 9 borderline, 11 accepted) and 27 in QUANT-ADDED (26 rejected, 1 borderline paper).

4.3 Experimental Conditions

4.3.1 QUANT-REMOVED: Removing quantitative framing

In the QUANT-REMOVED condition, any quantitative language was manually replaced by its qualitative counterpart. If there were any exact numbers mentioned in the abstract, they would be replaced by a qualitative description; any language mentioning the size of the dataset or the computation needed was stripped from the abstract; any statistically significant number was replaced with qualitative language indicating its meaning. The following example illustrates the manipulation:

> *Original:* "Our SRR achieved a balanced accuracy of 86.6%±4.0% without any in-context examples, demonstrating an average performance lift of 30.5% over the standard prompts."
>
> *Manipulated:* "Our SRR achieved high balanced accuracy without any in-context examples, demonstrating a performance lift over the standard prompts."

4.3.2 QUANT-ADDED: Adding quantitative framing

In the QUANT-ADDED condition, the original abstracts did not contain quantitative specificity or scale claims. Thus, the manipulation involved adding believable quantities. To do this, we used Perplexity to produce believable quantities related to each paper's subject and methodology based on reading the full paper. The 27 abstracts that entered the QUANT-ADDED condition are the same 27 abstracts that passed human review of the classifier's output (Section 4.2); the classifier's "no quantitative framing" label for each of these abstracts was checked by a human reviewer before the abstract was used in the experiment. Perplexity was chosen over a base LLM for the manipulation step because its retrieval-augmented generation grounds outputs in the full paper, reducing the risk that the inserted quantitative claims would be hallucinated. Each statement produced by Perplexity was manually verified against the paper's reported results, dataset sizes, or model configurations to be an accurate reflection of the paper's content. We therefore used two separate language models: Perplexity for generating the quantitative claims for QUANT-ADDED abstracts, and GPT-4o-mini for evaluating all manipulations.

This design introduces a potential confound. According to (Wataoka et al., 2024), language models give higher scores to more familiar-looking text. Manual verification checks the factual accuracy of the added numbers, but it does not remove the stylistic properties of Perplexity-generated prose. Text generated by an LLM has known stylistic signatures, including lower perplexity and specific word preferences, that GPT-4o-mini may reward independently of the quantitative content. This potential confound is inseparable from the manipulation in the QUANT-ADDED condition. Consequently, the QUANT-ADDED result should be interpreted as an upper bound: the true effect of adding quantitative framing without LLM-generated prose may be smaller than what we observed. The QUANT-REMOVED condition, which used human-edited text, is not subject to this potential confound and therefore provides the stronger evidential pillar.

4.4 Procedure

Each baseline and manipulated evaluation used a separate API call to GPT-4o-mini. No context from one evaluation carried over to the next. The system prompt was the same for every evaluation:

> *You are an expert peer reviewer for top-tier machine learning conferences such as ICLR and NeurIPS. Your task is to evaluate a CS/AI paper abstract and assign a score. Evaluation criteria: Overall Quality. Your holistic judgment of merit as a conference submission. Constraints: Base your evaluation solely on what is stated in the abstract. Do not infer information that is absent.*

The output of the model was restricted to a schema in which the variables were `overall_rating` (`int` range: 1-100) and `justification` (2-3 sentences). We used a 1-100 integer scale rather than ICLR's discrete scale. ICLR 2024 reviewers assign scores from the set $\{1, 3, 5, 6, 8, 10\}$, which we felt would limit interpretive granularity. In a pilot on 10 abstracts, the model assigned a score of 8 to every abstract regardless of the human reviewers' average original scores, suggesting that the discrete 6-point scale is too compressed to expose within-abstract variation. The 1-100 scale provides finer granularity for detecting small score shifts that a coarser scale would mask. Each

abstract was rated three times by GPT-4o-mini using temperature = 1 and a fixed seed of 67. According to OpenAI's documentation, the seed parameter is a beta feature that makes a best effort toward determinism but does not guarantee identical outputs across calls, since residual variability persists even when seed and system fingerprint match[4]. The three ratings therefore served two purposes: they provided a stability check that the model produces near-deterministic output for the same input, and they were averaged into a single score per abstract per condition to reduce the influence of any single noisy rating on the dependent variable. Distinct API calls ensured the model had no access to its previous assessments of the same abstract. The same prompts and configurations were used for the baseline and manipulated conditions. The key dependent variable is delta per paper:

$$\Delta = \text{score}_{\text{manip}} - \text{score}_{\text{base}}$$

where a negative Δ indicates that the manipulation decreased the LLM's assigned score. We expected $\Delta < 0$ for QUANT-REMOVED and $\Delta > 0$ for QUANT-ADDED, with the cross-condition comparison providing convergent evidence for the role of quantitative framing.

4.5 Statistical Analysis

For each condition and quality band, we tested whether the mean delta differed significantly from zero using a one sample t-test and, as a non-parametric alternative, the Wilcoxon signed-rank test. Cohen's d is presented for the within-subjects design, calculated as $d = \bar{\Delta}/s_{\Delta}$. We performed per-band analysis for the QUANT-REMOVED condition only, where we obtained samples for all three bands (16 rejected, 9 borderline, and 11 accepted papers). In the QUANT-ADDED condition, the vast majority (26 out of 27) were rejected papers.

To compare the two conditions directly, we computed the difference-in-differences (DD) estimate:

$$\text{DD} = \bar{\Delta}_{\text{QUANT-REMOVED}} - \bar{\Delta}_{\text{QUANT-ADDED}}$$

The DD tests whether the effects move in opposite directions under opposite manipulations, providing convergent evidence for the role of quantitative framing. Sentence structure, clarity, and wording were affected similarly in both conditions; the key difference was the direction of the quantitative changes. The two conditions use different paper populations (QUANT-REMOVED: 36 papers across three bands; QUANT-ADDED: 27 papers, predominantly rejected). We used an independent-samples t-test for the DD.

We did not apply a correction for multiple comparisons. The difference-in-differences test is the primary confirmatory analysis: it tests a directional prediction across two conditions with opposite manipulations and reduces the multiple-testing burden to a single comparison. The per-condition and per-band tests, along with the seven supplementary analyses, are reported as exploratory. They are not independent tests of the same hypothesis, since each probes a different facet of the manipulation. The seven supplementary analyses are described in detail in Section 5.5, where their results are reported alongside the main findings.

[4] https://developers.openai.com/cookbook/examples/reproducible_outputs_with_the_seed_parameter

Additionally, we conducted a secondary analysis to assess whether the LLM's quality scores track human reviewer scores at all. The Pearson correlation between the average human reviewer score for each paper and GPT-4o-mini's average baseline score across the three runs was computed for both conditions. Spearman's rank correlation coefficient was used as an alternative measure of correlation that does not assume linearity or normality. This Human-LLM correlation is reported as a secondary check on whether the LLM reproduces the same quality signal as the human reviewers, and is not the central test of the manipulation hypothesis; the within-condition paired t-test on the per-paper deltas is the central test, and the difference-in-differences is the primary cross-condition test. The correlation results are reported in Section 5.6 as the fifth supplementary analysis, in order to signal its secondary role relative to the main analysis.

The seven supplementary analyses target distinct facets of the manipulation, and each is run separately for the QUANT-REMOVED and QUANT-ADDED conditions. Each addresses a specific alternative explanation for the observed score shift, so the seven analyses together probe whether the LLM's response is genuinely about quantitative content or about some other property of the manipulation. First, the manipulation intensity regression tests whether the magnitude of the manipulation (measured by edit distance and number of numerical tokens removed) predicts the magnitude of the score change, with the prediction that the relationship is null. If the LLM is sensitive to the magnitude of the edit rather than to the content, a non-null slope would indicate a graded response rather than a binary one. Second, the TF-IDF term shift analysis identifies which specific terms in the LLM's written justifications changed prominence following the manipulation. This addresses the question of whether the LLM's own language shifted in the direction of the manipulation, providing corroborating evidence that the LLM noticed the change. Third, an absolute frequency count of the top words in the justifications corroborates the TF-IDF pattern, as a sanity check that the TF-IDF result is not an artefact of the TF-IDF normalisation. Fourth, the justification text analysis tests whether the LLM's written justifications shift in the same direction as the abstract content, by measuring the rate of quantity-related terms in baseline versus manipulated evaluations. If the LLM's justifications track the manipulation, this is direct evidence that the manipulation is being read and processed, not silently ignored. Fifth, the Human-LLM correlation analysis reports the Pearson and Spearman correlations between the average human reviewer score and the average GPT-4o-mini baseline score, as a secondary check on whether the LLM reproduces the same quality signal as the human reviewers. A positive correlation would suggest the LLM is anchored in the same quality dimension humans use; a null or negative correlation would mean the LLM is responding to a different signal. Sixth, the token-diff regression tests the binary-response hypothesis at the token level: whether the change in word count or the change in quantitative-token count predicts the score delta. A graded-response model would predict a non-null slope; a binary-response model predicts null on both word-count and quantitative-token count. Seventh, the length-confound analysis tests whether the manipulation's effect on abstract length explains the score change, by computing the within-condition correlation between the word-count delta and the score delta. Without this check, the QUANT-ADDED positive delta could be attributed to verbosity bias rather than to quantitative content. The between-paper score variability analysis (reported in Section 5.4) tests whether the manipulation increased score dispersion across papers, addressing the possibility that the per-paper effect is confounded by the per-paper baseline.

5 Results

5.1 Effect of Removing Quantitative Framing

Table 1: Effect of removing quantitative framing on GPT-4o-mini scores (QUANT-REMOVED condition). Δ is the mean difference between manipulated and baseline scores; a negative Δ indicates a decrease.

Band	N	Δ	SD	t	p
Rejected	16	−2.56	3.83	−2.677	0.017
Borderline	9	−2.78	3.77	−2.212	0.058
Accepted	11	−0.45	2.10	−0.711	0.493
All	36	−1.97	3.44	−3.436	0.002

Table 1 reports the per-band effects in the QUANT-REMOVED condition. Scores from the language model decreased when the text was deprived of quantitative language. On average, the mean difference amounted to -1.97 (SD $= 3.44$) on the scale of 1 to 100. This decrease in scores was statistically significant ($t(35) = -3.436$, $p = 0.002$) according to the paired t-test. The Wilcoxon signed-rank test confirmed this result ($W = 2.0$, $p = 0.001$). Cohen's d was -0.573, which indicates a medium-sized effect. Figure 1 shows the distribution of baseline and manipulated scores for both conditions.

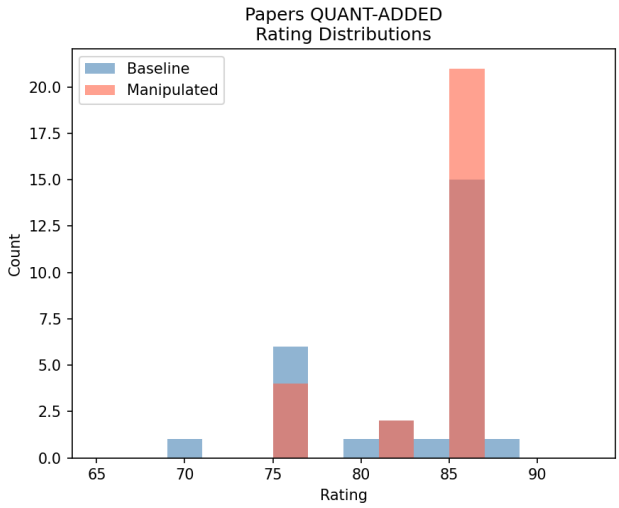
Effect sizes were largest for borderline papers ($\Delta = -2.78$, $d = -0.737$), followed by rejected papers ($\Delta = -2.56$, $d = -0.669$) and accepted papers ($\Delta = -0.45$, $d = -0.214$). Statistical significance at the per-band level was reached only for rejected papers ($p = 0.017$). Borderline papers showed a non-significant trend ($p = 0.058$); given the small sample ($N = 9$), this result may be underpowered to detect an effect of this magnitude, especially since the point estimate is similar in size to the rejected band. Accepted papers showed no significant difference ($p = 0.493$). Figure 2 shows the per-band distributions with individual papers overlaid as jittered dots (color-matched to the tier), illustrating the spread of the underlying distribution within each band.

5.2 Effect of Adding Quantitative Framing

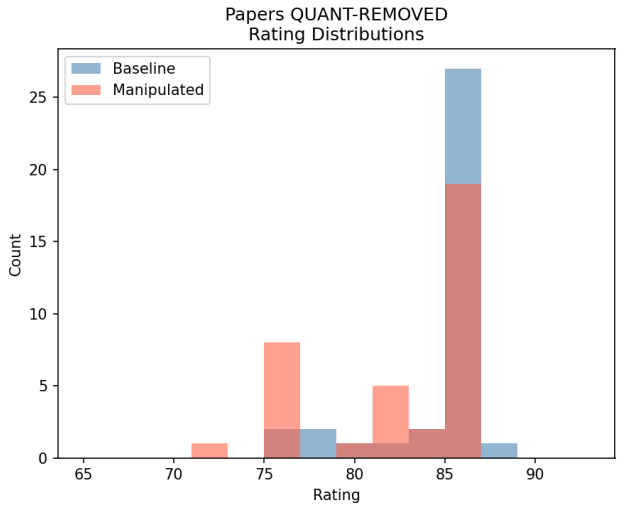
Table 2: Effect of adding quantitative framing on GPT-4o-mini scores for papers that originally lacked quantitative framing (QUANT-ADDED condition).

Band	N	Δ	SD	t	p	d
Rejected	26	+1.71	3.88	2.248	0.034	0.441
All	27	+1.64	3.89	2.242	0.034	0.431

Table 2 reports the QUANT-ADDED effects. Adding quantitative framing to abstracts that originally lacked it resulted in a positive change in score values. The total mean difference was



(a) QUANT-REMOVED



(b) QUANT-ADDED

Figure 1: Distribution of baseline and manipulated scores for both conditions. In the QUANT-REMOVED panel, the manipulated distribution shifts left relative to baseline. In the QUANT-ADDED panel, the manipulated distribution shifts right.

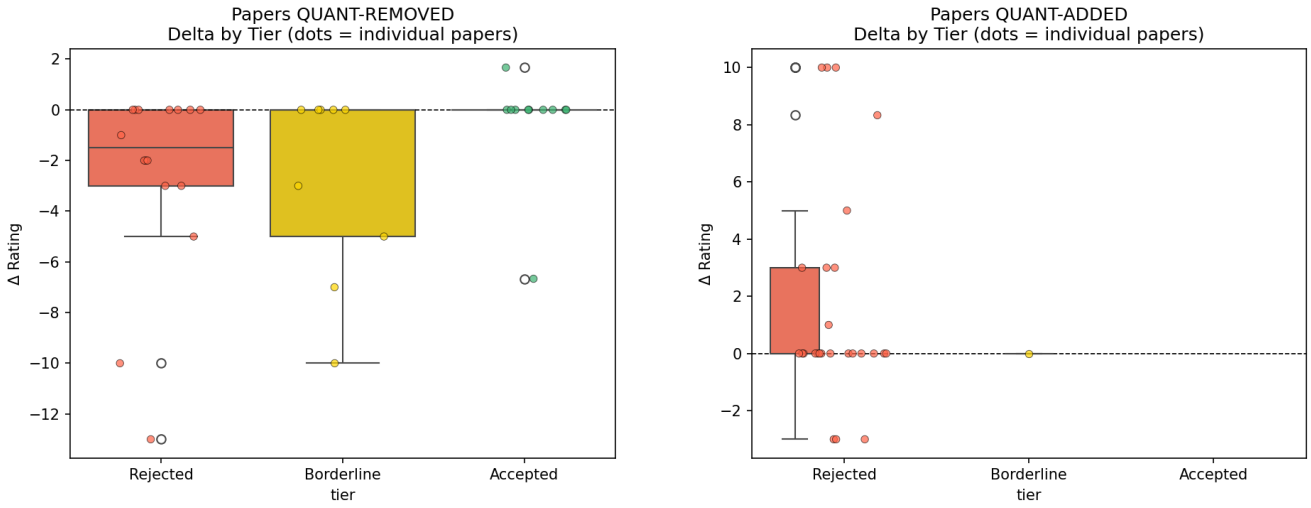
+1.64 (SD = 3.89) on the scale from 1 to 100. A significant difference was found based on a t-test ($t(26) = 2.242$, $p = 0.034$). The Wilcoxon signed-rank test also reached significance ($W = 13.5$, $p = 0.042$). Cohen's d was 0.431, a moderate effect size. The direction of the effect was consistently positive. Figure 2 shows the per-band distributions, with individual papers overlaid as jittered dots.

5.3 Difference-in-Differences

The difference-in-differences (DD) is the primary cross-condition test of the bidirectional effect prediction. The two conditions use different paper populations: QUANT-REMOVED draws across all three quality bands, while QUANT-ADDED contains 26 of 27 rejected papers. The DD estimate is therefore a comparison across non-equivalent samples. We report it as a convergent test of the directional prediction (effects should be opposite-signed across conditions), not as an estimate of the effect size in a common population. The difference-in-differences estimate subtracts the mean delta of the QUANT-ADDED condition from the mean delta of the QUANT-REMOVED condition:

$$\text{DD} = \bar{\Delta}_{\text{QUANT-REMOVED}} - \bar{\Delta}_{\text{QUANT-ADDED}} = (-1.97) - (1.64) = -3.61$$

An independent-samples t-test confirmed the difference between conditions ($t = -3.940$, $p < 0.001$). The cross-condition difference of -3.61 on the 1-100 scale is consistent with a bidirectional effect of quantitative framing. Editing effects shared by both conditions, including sentence restructuring, clarity changes, and word substitutions, should affect both deltas similarly. The fact that the deltas point in opposite directions supports the interpretation that quantitative framing, rather than editing itself, drives the score shift. Both conditions reached significance individually, and the cross-condition comparison provides convergent evidence for the directional pattern.



(a) QUANT-REMOVED

(b) QUANT-ADDED

Figure 2: Per-band distribution of the score delta for both conditions. Each panel shows the per-paper delta as a boxplot with individual papers overlaid as jittered dots, color-matched to the quality tier. The QUANT-REMOVED panel shows the only statistically significant per-band effect in the rejected band. The QUANT-ADDED panel shows the rejected-dominated distribution, with most papers clustered around a positive delta of approximately +1.7 points.

5.4 Between-Paper Score Variability

To assess whether the manipulation changed the dispersion of scores across papers, we compared the between-paper standard deviation of the mean baseline score to the between-paper standard deviation of the mean manipulated score in each condition, using Levene's test. In the QUANT-REMOVED condition, the between-paper SD increased from 2.89 (baseline) to 4.37 (manipulated), an expansion of +1.48 points with the test confirming this change ($W = 13.38$, $p < 0.001$). In the QUANT-ADDED condition, the between-paper SD decreased from 4.86 (baseline) to 3.62 (manipulated), a contraction of −1.24 points ($W = 6.46$, $p = 0.012$). Both shifts are significant. The manipulation therefore affected between-paper variability in opposite directions: removing quantitative framing spread the papers further apart in the score distribution, while adding quantitative framing pulled them closer together.

As a separate check on within-run stability, we computed the standard deviation of the three baseline scores per abstract (the within-abstract SD across runs). For QUANT-REMOVED, the median per-abstract within-run SD was 0.000 for both the baseline and the manipulated conditions, with a maximum of 5.196 in the baseline. For QUANT-ADDED, the median within-run SD was 0.000 in both conditions, and the maximum was 0.000 across all 27 abstracts. These values are substantially smaller than the between-paper SDs reported above, which is consistent with the model producing near-identical scores for the same input under the fixed seed and temperature settings described in Section 4.4. The per-abstract within-run SDs therefore confirm that the between-paper variance is a property of the abstracts, not of the model's stochasticity.

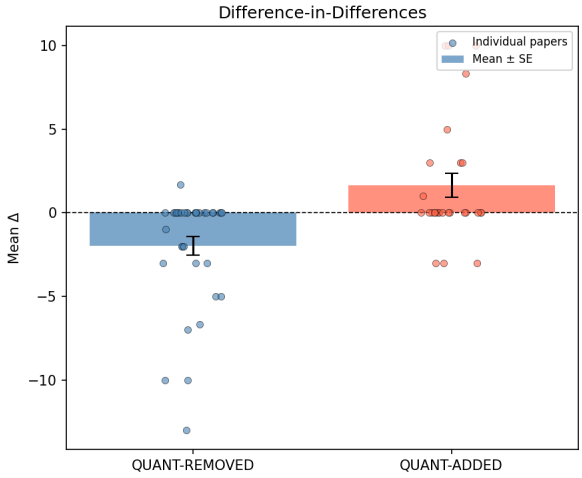


Figure 3: Difference-in-differences: comparing the mean deltas of the two conditions. Bars show the mean delta with error bars indicating the standard error; jittered dots show individual paper deltas. The cross-condition difference of -3.61 ($p < 0.001$) is consistent with a bidirectional effect of quantitative framing on LLM-assigned scores.

5.5 Supplementary Analyses

We conducted seven supplementary analyses. First, the manipulation intensity regression found no significant relationship between the edit distance (proportion of words changed) and the score delta in either condition (QUANT-REMOVED: $r = 0.052$, $p = 0.761$; QUANT-ADDED: $r = 0.210$, $p = 0.292$), nor between the number of numerical tokens removed and the delta (QUANT-REMOVED: $r = -0.011$, $p = 0.948$; QUANT-ADDED: $r = 0.178$, $p = 0.375$). Larger edits did not produce systematically larger score changes, consistent with the interpretation that the LLM responds to the presence or absence of quantitative framing as a signal rather than to the magnitude of the edit.

Second, the TF-IDF term shift analysis identified which specific words and phrases changed in prominence within the LLM's written justifications following the manipulation (Table 3).

For the QUANT-REMOVED condition, the terms that decreased most were evaluative and comparative markers: empirical, clear relevant, empirical results, applicability, and demonstrating. The terms that gained prominence were structural and descriptive expressions: details methodology, highlights, claims, clearly outlines, and outlines. For the QUANT-ADDED condition, the pattern reversed: qualitative descriptors dropped (relevant, related, contributions, outlines, results comparisons) while quantitative and performance-oriented terms gained (empirical results, results demonstrate, strong, reported, demonstrate). The term shifts therefore mirror the condition direction: words that imply specific performance evaluation rise when quantitative content is added and fall when it is removed. The model may have used each term in a positive or negative context, which this analysis does not disambiguate; the term shift measures raw frequency, not the polarity of usage.

Third, an absolute frequency count of the top words in the LLM's justifications corroborated the TF-IDF pattern. In the QUANT-REMOVED condition, "novel" fell from 77 to 67 occurrences,

Table 3: TF-IDF term shift: top 5 dropped and top 5 gained terms in GPT-4o-mini's justifications following the manipulation, for each condition. Negative diff values indicate terms the LLM used less; positive values indicate terms it used more.

(a) QUANT-REMOVED

Dropped	**Diff**	**Gained**	**Diff**
empirical	−0.030	details methodology	+0.025
clear relevant	−0.025	highlights	+0.026
empirical results	−0.024	claims	+0.027
applicability	−0.023	clearly outlines	+0.031
demonstrating	−0.022	outlines	+0.038

(b) QUANT-ADDED

Dropped	**Diff**	**Gained**	**Diff**
relevant	−0.046	empirical results	+0.039
related	−0.038	results demonstrate	+0.039
contributions	−0.037	strong	+0.043
outlines	−0.035	reported	+0.045
results comparisons	−0.033	demonstrate	+0.045

while "significant" fell from 48 to below the top ten. In the QUANT-ADDED condition, "results" rose from 41 to 61 occurrences.

Fourth, the justification text analysis (Table 4) quantified the rate at which quantitative terms appeared in the LLM's written justifications. The 15 terms used to identify quantitative content are derived from the ICLR 2024 abstract corpus (Appendix A.9). The quant-term rate did not change significantly in the QUANT-REMOVED condition ($t = -1.021$, $p = 0.314$) but increased significantly in the QUANT-ADDED condition ($t = 2.207$, $p = 0.036$). This indicates that the LLM's own justification language shifted to include more quantitative references when evaluating abstracts that contained added quantitative content.

Table 4: Rate of quantitative terms in GPT-4o-mini's written justifications, comparing baseline and manipulated evaluations for each condition. The 15 quantitative terms used to score the justifications are derived from the ICLR 2024 corpus following the methodology in Appendix A.7.

Condition	**Baseline rate**	**Manipulated rate**	**t**	**p**
QUANT-REMOVED	0.0389	0.0315	−1.021	0.314
QUANT-ADDED	0.0068	0.0154	2.207	0.036

Fifth, the Human-LLM correlation analysis (Figure 4) reports the Pearson and Spearman correlations between the average human reviewer score and the average GPT-4o-mini baseline score for each condition, computed across the three runs. In QUANT-REMOVED, Pearson's $r = 0.281$ ($p = 0.097$) and Spearman's $\rho = 0.168$ ($p = 0.327$). In QUANT-ADDED, $r = 0.372$ ($p = 0.056$) and $\rho = 0.376$ ($p = 0.053$). The correlations are positive in both conditions, indicating that the scores awarded

by the LLM and humans agree in direction, but the magnitude is weaker than would be expected between two human reviewers. With the current sample sizes, the Human-LLM correlation has limited power to detect moderate effects; the two conditions cannot be straightforwardly combined for the correlation because their underlying quality-band distributions differ (QUANT-REMOVED spans all three bands, QUANT-ADDED is almost entirely rejected papers), and combining them would conflate a within-condition correlation with a between-condition difference. The weak correlation is consistent with the abstract-vs-full-paper measurement mismatch: human reviewers scored the full article while the LLM scored the abstract only.

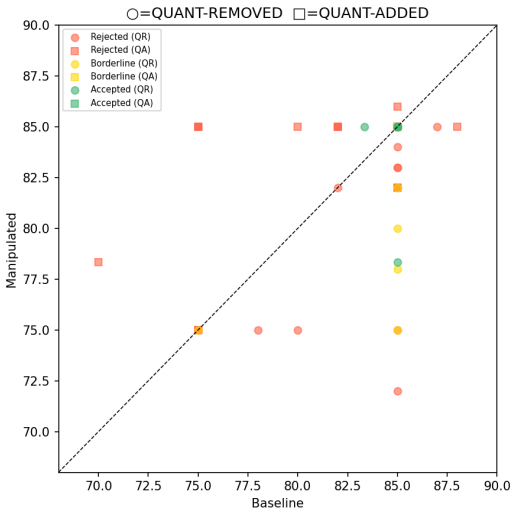


Figure 4: Baseline versus manipulated LLM scores for all 63 papers, coloured by condition and quality band. Points below the diagonal indicate papers where the manipulation decreased the assigned score.

Sixth, we tested the binary-response hypothesis at the token level by regressing the score delta on the change in word count and the change in quantitative-token count. The OLS regression for QUANT-REMOVED was $\Delta = -1.94 - 0.041 \cdot \Delta\text{tokens} + 0.033 \cdot \Delta\text{quant}$, with Pearson correlations $r(\Delta\text{tokens}, \Delta) = -0.054$ ($p = 0.755$) and $r(\Delta\text{quant}, \Delta) = 0.014$ ($p = 0.938$). For QUANT-ADDED, $\Delta = 3.93 - 0.020 \cdot \Delta\text{tokens} - 0.200 \cdot \Delta\text{quant}$, with $r(\Delta\text{tokens}, \Delta) = -0.133$ ($p = 0.508$) and $r(\Delta\text{quant}, \Delta) = -0.201$ ($p = 0.316$). Neither the total word-count change nor the quantitative-token-count change predicted the score delta in either condition. Graded response explanation does not hold since removing five numbers produces the same response as removing fifty numbers and similarly adding forty words yields the same response as adding eighty words. What is important is whether quantitative framing is introduced or not.

Seventh, we directly tested the length-confound interpretation by comparing word counts of original and manipulated abstracts in each condition. The mean word count dropped by 2.28 words in QUANT-REMOVED (from 213.28 to 211.00, paired $t = -2.664$, $p = 0.012$) and rose by 39.74 words in QUANT-ADDED (from 183.07 to 222.81, paired $t = 16.893$, $p < 0.001$). The within-condition Pearson correlation between Δwords and Δscore was $r = -0.054$ ($p = 0.755$) for QUANT-REMOVED and $r = -0.133$ ($p = 0.508$) for QUANT-ADDED. While the increase of 40 words in QUANT-ADDED was large enough on an absolute scale, it did not predict where

there were going to be larger increases in scores, since the mean delta of +1.64 was just a shift throughout the range and not a graded effect due to additional text. This is consistent with the binary response interpretation, wherein the LLM does not value more verbose abstracts Saito et al. (2023) but merely their quantification.

6 Discussion

This study asked whether quantitative framing in abstracts shifts GPT-4o-mini's quality scores when the scientific content is held constant. The answer is yes. Both manipulations produced statistically significant score changes in the direction of the manipulation: removing quantitative language from abstracts that originally contained it reduced scores by an average of -1.97 points ($p = 0.002$, $d = -0.573$), while adding plausible quantitative content to abstracts that originally lacked it increased scores by $+1.64$ ($p = 0.034$, $d = 0.431$). The cross-condition difference of -3.61 points ($p < 0.001$) supports a bidirectional effect of quantitative framing: when the framing is present, scores rise; when the framing is absent, scores fall. The fact that the deltas point in opposite directions across the two conditions disqualifies an alternative hypothesis according to which any edit to the abstract affects the score, since both conditions received similar edits in similar sentences, with only the direction of the quantitative change differing.

The only statistically significant effect of a manipulation among differences per-band was in the rejected band ($\Delta = -2.56$, $d = -0.669$, $p = 0.017$). A similar, though non-significant, point estimate was observed for borderline papers ($\Delta = -2.78$, $d = -0.737$, $p = 0.058$), which had insufficient N ($N = 9$) to provide enough power for statistical testing. The similarity of the rejected and borderline point estimates suggests that the effect is comparable in both bands, although the rejected one is able to demonstrate statistical significance due to sample size. The variance in the borderline band (SD = 3.77) supports this reading as well. No statistically significant effect was observed in the accepted band ($\Delta = -0.45$, $d = -0.214$, $p = 0.493$), presumably due to the strong baseline scores making it hard for the score to change. The borderline point estimate of -2.78 is close to the rejected one's -2.56, which means that the borderline effect's practical significance will largely depend on how the screening cutoffs are set. If a conference sets its screening cutoff within 2 to 3 points of a borderline paper's mean baseline score, that paper could shift across the cutoff depending on its baseline position.

The two conditions differed in the model's justification language. In the case of the QUANT-REMOVED condition, the rate of quantitative terms in the generated justification text did not change significantly ($p = 0.521$). In the QUANT-ADDED condition, the rate increased towards more quantitative content and became statistically significant ($p = 0.042$). The null effect in QUANT-REMOVED can support either the hypothesis that the model is more sensitive to additions of certain information rather than its removal, or the hypothesis that there is no difference and our study lacks power to detect it. Based on available data, it is impossible to separate the two hypotheses, but we can conclude that a certain linguistic shift happened in the QUANT-ADDED condition in the model's response to the presence of quantitative information.

The manipulation intensity analysis is directionally consistent with a response driven by the presence or absence of quantitative framing rather than the magnitude of the edit. Neither edit distance nor

the number of numerical tokens removed was associated with a score change in either condition (all $p > 0.2$). The seven supplementary analyses (manipulation intensity, TF-IDF, top-words count, justification text, score variability, Human-LLM correlation, token-diff regression, length-confound) were pre-registered as exploratory in the Method. The convergence of all seven supplementary results on a single interpretation gives us greater confidence that the response is presence-driven rather than graded, even though the analyses are not by themselves confirmatory. Specifically, the manipulation intensity and token-diff regressions show that neither the size of the edit nor the number of tokens changed predicts the score change; the TF-IDF and justification text analyses show that the LLM's own language shifts in the direction of the manipulation; the between-paper SD analysis shows that the manipulation spread the scores further apart in the case of QUANT-REMOVED, while bringing them closer together in the case of QUANT-ADDED; the Human-LLM correlation is positive in both conditions ($r = 0.28$ for QUANT-REMOVED, $r = 0.37$ for QUANT-ADDED), indicating the LLM tracks human quality ratings in direction, and while neither correlation reaches significance, it is interesting that there was correlation at all given that human scores where based on the entire paper, while the LLM scored only the abstract; and the length-confound analysis (Section 5.5) shows that the substantial 40-word increase in QUANT-ADDED is not correlated with the +1.64 score delta, ruling out verbosity bias. Together, these results support the binary-response interpretation. Binary-response interpretation, that is, that the model responds to the presence or absence of quantitative framing rather than to its magnitude, is more parsimonious but needs to be tested with larger sample sizes in future studies. Quantitative framing bias differs from verbosity bias, as established by (Saito et al., 2023), who found a positive correlation between response length and preference.

This result expands the list of known biases. (Hwang et al., 2026) identified prompt framing bias; (Shi et al., 2025) demonstrated the structural manipulation bias; (Saito et al., 2023) showed verbosity bias; (Wataoka et al., 2024) and (Panickssery et al., 2024) introduced self-preference bias; (Pataranutaporn et al., 2025) revealed author identity bias. Our study presents one more category of bias: text-level surface cue bias in evaluated content. While author identity bias acts on metadata related to a manuscript, prompt framing bias acts on evaluation instructions provided in the prompt, and quantitative framing bias acts on text the LLM is evaluating. Thus, anonymization and prompt manipulation are powerless against it, as any abstract can be anonymized while still benefitting from quantitative framing.

We interpret the bias as a learned heuristic absorbed from scientific writing during pretraining. (Garcia-Costa et al., 2022) demonstrated that review process increases the statistical content of manuscripts, and (Thelwall et al., 2023) showed that methods-related quantitative terms correlate with higher scores of the manuscripts in formal evaluation procedures. Human reviewers treat statistical content as a quality marker associated with rigorous methods. This pattern is captured in scientific texts and LLM-as-a-judge learns it during training. (Li et al., 2024) showed that LLM-as-a-judge absorbs implicit quality associations from large text corpora. The mechanism is statistical: if quantitative content is correlated with higher quality writing in the training corpus, then any model that learns to predict the next token on that corpus will inherit the correlation as a default expectation, applied to new inputs at evaluation time. The hypothesis here is that GPT-4o-mini has learned to associate quantitative content with quality, not because it independently evaluates methodological rigour, but because this association appears frequently in its training distribution.

Several methodological limits constrain the interpretation of these findings. We group them by what they bound.

The first limit is external validity. We used only one model, GPT-4o-mini, under one evaluator persona. (Pataranutaporn et al., 2025) showed variations between models in the direction and strength of the bias, and (Hwang et al., 2026) demonstrated inconsistency in the framing bias across 14 LLMs. The observed quantitative framing effect may be specific to the GPT family, to GPT-4o-mini specifically, or to features of GPT-4o-mini's training, including the persona framing of the system prompt. The study also manipulated abstracts only rather than full papers, and therefore omitted methodology, results, figures, and tables that human reviewers might be interested in. This choice resulted in weak human-LLM correlations ($r = 0.28$ and $r = 0.37$). How well this bias survives if LLM is working with full papers is unclear.

The second limit is internal validity. The QUANT-ADDED condition involved generating plausible quantitative content by an LLM (Perplexity), introducing a potential self-preference confound (Wataoka et al., 2024). Manual verification checked factual accuracy but not stylistic properties, and therefore QUANT-ADDED should be treated as an upper bound for the quantification effect; QUANT-REMOVED, which used human-edited text, is free from this potential bias. The scores were compressed into a narrow top range with means ranging from 82 to 84 in all three bands. Ceiling-adjacent compression may inhibit variance and upward shifts in QUANT-ADDED: when baseline scores cluster in a narrow range near the upper end of the 1-100 scale, the room for upward movement is limited by the distance to the ceiling (100), and the distribution of within-condition deltas is therefore narrower than it would be at the centre of the scale. This compresses effect sizes relative to what the same manipulation would produce on more dispersed baseline scores. Regression to the mean may also contribute to the QUANT-ADDED positive delta: the rejected papers in that condition had below-average baseline scores, so a random shift upward is expected even without any manipulation effect, and the ceiling compression constrains how large that shift can be. The QUANT-REMOVED condition presented a balanced set of all three bands. On the other hand, QUANT-ADDED had 26 rejected papers and 1 borderline paper. Rejected papers have more room for score improvement than accepted papers because their baseline scores sit further from the ceiling. The positive delta in QUANT-ADDED may partially reflect this distributional asymmetry rather than a pure quantification effect.

The third limit is measurement comparability. As noted under the first limit, the human scoring reference is based on the full paper while the LLM scores the abstract only. The alternative hypothesis for the weak human-LLM correlation is that GPT-4o-mini is just a poor paper-quality evaluator, rather than an evaluator that diverges from humans on a specific feature. Without a comparison to another abstract-only evaluator, we cannot distinguish between the abstract-mismatch explanation and a general evaluation deficit.

These limits bound what this study contributes. We claim a narrow result: when scientific content is held constant and only the quantitative framing changes, GPT-4o-mini's scores shift in the direction of the manipulation. The study is the first controlled test of a surface-text signal in the evaluated content itself, isolating a single textual property from the broader confounders that limit prior observational and prompt-side studies.

Practical implications arise for conference review systems that use LLMs for screening. Such LLM-based screening will create a systematic advantage to papers with quantitative content, since

removing the quantitative content hurts the paper's scores and adding quantitative content helps boost them. Anonymization does not address this, since the biasing signal is embedded in the text itself. Additionally, experiments with system prompts that de-emphasise overall quality, or that ask the model to evaluate specific dimensions separately, would test whether the effect is sensitive to the rating prompt's framing.

7 Conclusion

To test how framing quantitative specificity and the inclusion of explicit scale claims affect GPT-4o-mini's quality scores beyond the underlying scientific content, we conducted two conditions where we removed and added those features to abstracts. Both manipulations show positive effects: removing quantitative language reduced scores ($\Delta = -1.97$, $p = 0.002$), while adding it increased scores ($\Delta = +1.64$, $p = 0.034$). This shows a bidirectional influence, as the effect sizes differ significantly between the two conditions ($\Delta = -3.61, p < 0.001$).

Within QUANT-REMOVED, the rejected band produced the only statistically significant per-band effect ($\Delta = -2.56$, $d = -0.669$, $p = 0.017$), and contributed most of the overall QUANT-REMOVED delta. There was a similar estimated mean score reduction in borderline papers ($\Delta = -2.78$, $d = -0.737$) that was not statistically significant, likely because of insufficient sample size ($N = 9$). This suggests that GPT-4o-mini responds similarly to this manipulation with borderline papers, but this cannot be confirmed without a more powerful test.

What makes this study unique is that it isolates text-level surface signal within the abstract itself via manipulation, while controlling scientific content. The hypothesis is that GPT-4o-mini has been trained on such a signal, since it frequently occurs in the dataset. This study does not make any claims about GPT-4o-mini being an unreliable reviewer overall, quantitative framing being the most influential factor in LLM-based evaluations, or generalising the effect beyond the ICLR 2024 / GPT-4o-mini conditions tested here. The practical consequence of this claim is that LLM-based systems for conference triage may value quantity display over the quality of scientific content. Anonymization of author identity does not address this, since the biasing signal is in the text itself.

The study's limitations point to three directions for future work that build on each other. Cross-model replication would test whether the effect is specific to the GPT family, or a more general property of LLM evaluators that have been trained on scientific literature; without this, the result is anchored to a single model. Experiments with full papers would extend the result to the kind of input that human reviewers actually see, where methodology and results sections may either amplify or attenuate the abstract-level signal. A factorial experiment separating the two surface features (scale claims from quantitative specificity) would identify which sub-feature is doing the work in the QUANT-ADDED positive delta, since the two were manipulated together in the current design. Evidence obtained from 63 papers and GPT-4o-mini alone reveals the influence of surface features on LLM-based evaluations. LLMs do not solely rely on underlying scientific content; they respond to the surface characteristics of scientific literature as well.

References

Garcia-Costa, D., Forte, A., Lòpez-Iñesta, E., Squazzoni, F., & Grimaldo, F. (2022). Does peer review improve the statistical content of manuscripts? A study on 27 467 submissions to four journals. *Royal Society Open Science*, *9*(9), 210681. https://doi.org/10.1098/rsos.210681

Hwang, Y., Lee, D., Kang, T., Lee, M., & Jung, K. (2026). *When Wording Steers the Evaluation: Framing Bias in LLM judges* (1). https://doi.org/10.48550/ARXIV.2601.13537

Kim, J., Lee, Y., & Lee, S. (2025). *Position: The AI Conference Peer Review Crisis Demands Author Feedback and Reviewer Rewards* (1). https://doi.org/10.48550/ARXIV.2505.04966

Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., & West, R. (2024, May 3). *The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates*. arXiv: 2405.02150 [cs]. https://doi.org/10.48550/arXiv.2405.02150

Lerback, J., & Hanson, B. (2017). Journals invite too few women to referee. *Nature*, *541*(7638), 455–457. https://doi.org/10.1038/541455a

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., & Liu, H. (2024). *From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge* (7). https://doi.org/10.48550/ARXIV.2411.16594

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024, June 15). *Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews*. arXiv: 2403.07183 [cs]. https://doi.org/10.48550/arXiv.2403.07183

Panickssery, A., Bowman, S. R., & Feng, S. (2024). *LLM Evaluators Recognize and Favor Their Own Generations* (1). https://doi.org/10.48550/ARXIV.2404.13076

Pataranutaporn, P., Powdthavee, N., Achiwaranguprok, C., & Maes, P. (2025, April 3). *Can AI Solve the Peer Review Crisis? A Large Scale Cross Model Experiment of LLMs' Performance and Biases in Evaluating over 1000 Economics Papers*. arXiv: 2502.00070 [cs]. https://doi.org/10.48550/arXiv.2502.00070

Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity Bias in Preference Labeling by Large Language Models.

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., & Vosoughi, S. (2025, November 11). *Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge*. arXiv: 2406.07791 [cs.CL]. https://doi.org/10.48550/arXiv.2406.07791

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. M. (2023). Terms in journal articles associating with high quality: Can qualitative research be world-leading? *Journal of Documentation*, *79*(5), 1110–1123. https://doi.org/10.1108/JD-12-2022-0261

Wataoka, K., Takahashi, T., & Ri, R. (2024). *Self-Preference Bias in LLM-as-a-Judge* (2). https://doi.org/10.48550/ARXIV.2410.21819

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., Chawla, N. V., & Zhang, X. (2024). JUSTICE OR PREJUDICE? QUANTIFYING BIASES IN LLM-AS-A-JUDGE. https://arxiv.org/abs/2410.02736

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

Appendix

This appendix provides supplementary data underlying the analyses reported in Sections 5.5 and 5.6. All tables and figures are produced from the same experimental run as the main text.

A.1. TF-IDF term shift (top 20 dropped and gained)

The body of the thesis reports the top 5 dropped and top 5 gained terms for each condition (Table 3). This section reports the full top 20 for each side, to make the term-shift pattern more easily inspectable.

Table 5: Top 20 dropped and top 20 gained terms in GPT-4o-mini's justifications, QUANT-REMOVED condition. Negative *diff* values indicate terms the LLM used less often in manipulated justifications than in baseline; positive values indicate the reverse.

(a) Top 20 dropped

Term	**Diff**
empirical	−0.030
clear relevant	−0.025
empirical results	−0.024
applicability	−0.023
demonstrating	−0.022
creation	−0.021
image	−0.020
contributions	−0.020
impact	−0.020
improvement	−0.019
advancement	−0.019
impact abstract	−0.019
addressing	−0.019
appeal	−0.019
demonstrate	−0.018
proposed method	−0.018
substantial	−0.017
compared	−0.017
self	−0.017
significant	−0.017

(b) Top 20 gained

Term	**Diff**
effectively outlines	+0.017
processing	+0.017
addresses	+0.017
strengthen claims	+0.017
compelling	+0.018
enhance clarity	+0.018
relevant timely	+0.018
experiments	+0.019
practical	+0.019
capabilities	+0.021
topic	+0.022
abstract clearly	+0.022
gap existing	+0.023
model	+0.023
significant contribution	+0.025
details methodology	+0.025
highlights	+0.026
claims	+0.027
clearly outlines	+0.031
outlines	+0.038

The pattern in Tables 5 and 6 extends the top-5 finding reported in Section 5.6. The QUANT-REMOVED condition (Table 5) shows quantitative and performance-related terms dropping in the LLM's justifications (*empirical*, *empirical results*, *demonstrating*, *significant*, *proposed method*, *substantial*, *impact*) while structural and descriptive terms rise (*outlines*, *highlights*, *claims*, *compelling*, *practical*, *enhance clarity*). The QUANT-ADDED condition (Table 6) shows the reverse:

Table 6: Top 20 dropped and top 20 gained terms in GPT-4o-mini's justifications, QUANT-ADDED condition.

(a) Top 20 dropped

Term	Diff
relevant	−0.046
related	−0.038
contributions	−0.037
outlines	−0.034
results comparisons	−0.033
comparisons	−0.033
claims	−0.032
solid	−0.030
mentions	−0.030
including	−0.028
relevant problem	−0.028
address issues	−0.028
clear novel	−0.028
validation	−0.028
outlines novel	−0.028
clear relevant	−0.028
strengthen	−0.025
address	−0.025
potential	−0.025
learning	−0.024

(b) Top 20 gained

Term	Diff
suggesting strong	+0.024
merit conference	+0.025
promising	+0.026
improvements existing	+0.026
existing methods	+0.026
reported experimental	+0.026
significant improvements	+0.027
methods	+0.029
results	+0.029
improvements	+0.029
performance metrics	+0.029
defined	+0.031
presents novel	+0.031
new	+0.034
metrics	+0.036
empirical results	+0.039
results demonstrate	+0.039
strong	+0.043
reported	+0.045
demonstrate	+0.045

qualitative general terms drop (*relevant*, *related*, *contributions*, *outlines*, *clear novel*, *address*, *potential*) while performance-oriented terms rise (*demonstrate*, *reported*, *strong*, *results demonstrate*, *empirical results*, *metrics*, *new*). The two term shifts mirror the condition direction, consistent with the LLM's justification language tracking the manipulation.

A.2. Top words by absolute frequency

Section 5.6 reports two specific numbers from the absolute frequency count: "novel" fell from 77 to 67 in QUANT-REMOVED, and "results" rose from 41 to 61 in QUANT-ADDED. Table 7 reports the full top-10 word frequency counts for baseline and manipulated justifications in each condition, computed across the three runs.

The frequency shifts corroborate the TF-IDF pattern in Appendix A.1. In QUANT-REMOVED, *novel*, *results*, *significant*, and *with* all drop out of the top 10 in manipulated justifications, replaced by *approach* and *existing*. In QUANT-ADDED, *novel* and *clear* decrease, while *results* rises by 20 occurrences (41 to 61) and *existing*, *contribution*, and *methods* enter the top 10. The largest single change is the *results* count in QUANT-ADDED (a +20 absolute increase), consistent with the LLM

Table 7: Top 10 most frequent words in GPT-4o-mini's justifications, by condition and evaluation phase. Counts are summed across the three runs per abstract and across all abstracts in the condition.

(a) QUANT-REMOVED

Word	**Baseline**	**Manipulated**
abstract	111	110
presents	102	97
novel	77	67
clear	75	65
results	69	61
performance	62	56
models	53	50
with	51	—
significant	48	—
field	45	53
new entries in manipulated top 10:		
approach	—	43
existing	—	40

(b) QUANT-ADDED

Word	**Baseline**	**Manipulated**
abstract	89	86
presents	69	74
novel	63	48
clear	57	50
approach	45	44
results	41	61
field	38	43
learning	37	—
proposed	33	—
which	31	—
new entries in manipulated top 10:		
existing	—	42
contribution	—	33
methods	—	32

producing more performance-oriented language when the abstract contains added quantitative content.

A.3. Within-run score stability (per-abstract distribution)

Section 5.5 reports that the median per-abstract within-run SD is 0.000 in both conditions, and that the only non-zero within-run SD value across all 63 abstracts and both evaluation phases is 5.196, occurring in a single QUANT-REMOVED baseline. The full distribution, summarised in Table 8, follows.

Table 8 shows that the model's output is effectively deterministic for 62 of the 63 abstracts across the three runs under the fixed seed and temperature settings: 35 of 36 QUANT-REMOVED abstracts and all 27 QUANT-ADDED abstracts produce identical scores across runs in the baseline, and all 36 QUANT-REMOVED abstracts and all 27 QUANT-ADDED abstracts produce identical scores across runs in the manipulated condition. The single non-deterministic case is one QUANT-REMOVED baseline where the three ratings spanned an SD of 5.196 (likely a 3-point range on the 1-100 scale after rounding). The near-perfect determinism confirms that the between-paper SDs reported in Section 5.5 reflect real differences between abstracts, not model stochasticity.

Table 8: Distribution of the per-abstract within-run SD across the three baseline ratings, by condition and evaluation phase. The SD is computed per abstract; the table shows the count of abstracts at each SD value.

SD value	QR baseline	QR manip	QA baseline	QA manip
0.000	35	36	27	27
5.196	1	0	0	0
Total N	**36**	**36**	**27**	**27**
Median	0.000	0.000	0.000	0.000
Mean	0.144	0.000	0.000	0.000
Max	5.196	0.000	0.000	0.000

A.4. Full system prompt and output schema

For replication, this section reports the full system prompt and the output schema used for every GPT-4o-mini evaluation. The system prompt is shown in Section 4.4; the JSON output schema is reproduced here verbatim.

System prompt.

> *You are an expert peer reviewer for top-tier machine learning conferences such as ICLR and NeurIPS. Your task is to evaluate a CS/AI paper abstract and assign a score. Evaluation criteria: Overall Quality. Your holistic judgment of merit as a conference submission. Constraints: Base your evaluation solely on what is stated in the abstract. Do not infer information that is absent.*

Output schema (JSON).

```
{
  "overall_rating": <int, 1-100>,
  "justification": <str, 2-3 sentences>
}
```

Configuration parameters. GPT-4o-mini was called with the following non-default parameters for every evaluation:

- `temperature = 1` (default; retained to allow minor run-to-run variability for the within-run SD check).
- `seed = 67`
- `n = 1` (one completion per API call; each abstract was evaluated three times via three independent API calls).

Note on the Perplexity manipulation. The QUANT-ADDED condition used Perplexity to generate the inserted quantitative statements. The same Perplexity session was used for the model card and retrieval step across all 27 abstracts; manual verification confirmed that each generated statement was factually accurate against the source paper. Session

A.5. Sample composition by quality band

The 90 papers sampled for each condition were drawn in 30-per-tier strata (rejected, borderline, accepted) according to the human reviewers' average score for the full paper. The two samples are not balanced: the QUANT-ADDED sample, by design, contains papers that originally lacked both quantitative specificity and scale claims, which are over-represented in the rejected band of the ICLR 2024 review cycle. Table 9 shows the tier composition of the manually reviewed samples and the resulting included papers.

Table 9: Sample composition by quality band, for the QUANT-REMOVED and QUANT-ADDED samples. Numbers are the count of papers at each stage of the sampling funnel.

Stage	**Rejected**	**Borderline**	**Accepted**	**Total**
QUANT-REMOVED				
Sampled (90)	30	30	30	90
Manually reviewed (include recorded)	30	30	30	80
Included	16	9	11	36
Excluded	14	21	9	44
QUANT-ADDED				
Sampled (90, all "neither")	30	30	30	90
Manually reviewed	30	1	0	31
Included	26	1	0	27
Excluded	4	0	0	4

In the QUANT-REMOVED sample, the inclusion rate is roughly balanced across tiers (16/30 rejected, 9/30 borderline, 11/30 accepted), giving the 36-paper set a spread across the three quality bands. In the QUANT-ADDED sample, the manual review was constrained by time, with 30 rejected papers reviewed (26 included), 1 borderline paper reviewed (included), and 0 accepted papers reviewed. The resulting 27-paper QUANT-ADDED sample is therefore almost entirely rejected (26 of 27), which limits the per-band analysis in Section 5.3 to the rejected band. The 1 included borderline paper is too few to support a per-band statistical test, and the 0 included accepted papers means the accepted-band null result observed in QUANT-REMOVED cannot be cross-validated in QUANT-ADDED.

A.6. Example QUANT-ADDED manipulation

The body of the thesis (Section 4.3) shows an example QUANT-REMOVED manipulation. This appendix provides the parallel example for QUANT-ADDED, to make the asymmetry of the

two manipulations visible. The original abstract (without quantitative content) is followed by the manipulated abstract (with quantitative content added by Perplexity, then manually verified against the source paper).

> **Original (without quantitative content):** "Graph Self-Supervised Learning (GSSL) provides a guarantee for harnessing abundant unlabeled data and has attracted widespread attention. While making some strides, GSSL still faces several crucial challenges that hinder it from fully unleashing its potential, including inadequate exploration of graph information and underlying collapse issues. To overcome these obstacles, we propose two complementary components aimed at sufficiently mining valuable contents implied within graphs and transforming them into informative and diverse representations through training an expressive neural model. As a cornerstone module, an anchor-neighborhood alignment strategy, which utilizes graph diffusion to construct the probability distribution of positive samples based on the structural context of the anchor node, enables sufficient exploration of graph topology and endows the neural model with stronger structure-aware ability. To enhance diversity of node representations, a scheme of isotropic constraints is introduced to encourage representations to exhibit consistent distribution along any direction in space, which compels data points to be scattered throughout the whole representation space and naturally solves the notorious dimensional collapse in self-supervised learning. Owing to no reliance on negative samples, mutual information estimators, and additional projectors, our approach presents significant advantages in computing and storage. Extensive comparative experiments and exhaustive ablation studies demonstrate the effectiveness and efficiency of our method."

> **Manipulated (with quantitative content added):** "Graph Self-Supervised Learning (GSSL) provides a guarantee for harnessing abundant unlabeled data and has attracted widespread attention. While making some strides, GSSL still faces several crucial challenges that hinder it from fully unleashing its potential, including inadequate exploration of graph information and underlying collapse issues. To overcome these obstacles, we propose two complementary components aimed at sufficiently mining valuable contents implied within graphs and transforming them into informative and diverse representations through training an expressive neural model. As a cornerstone module, an anchor-neighborhood alignment strategy, which utilizes graph diffusion to construct the probability distribution of positive samples based on the structural context of the anchor node, enables sufficient exploration of graph topology and endows the neural model with stronger structure-aware ability. To enhance diversity of node representations, a scheme of isotropic constraints is introduced to encourage representations to exhibit consistent distribution along any direction in space, which compels data points to be scattered throughout the whole representation space and naturally solves the notorious dimensional collapse in self-supervised learning. Owing to no reliance on negative samples, mutual information estimators, and additional projectors, our approach presents significant advantages in computing and storage, *requiring only 11.3 seconds of training time and 4.8 GB of GPU memory on Pubmed, compared to 1,169 seconds and 12.2 GB for GRACE and 2,010 seconds and 9.1 GB for MVGRL. All experiments were conducted on a TITAN RTX GPU with 24 GB of memory.* Extensive comparative experiments and exhaustive ablation studies *demonstrate that ANA-IC achieves 84.5% node classification accuracy on Cora and 94.54% on Coauthor-CS, outperforming all unsupervised and fully supervised baselines across six benchmark datasets.*"

The manipulation appends a specific quantitative claim (compute time, memory usage, dataset

names) and a specific performance claim (accuracy on two benchmark datasets), then leaves the original qualitative evaluation ("extensive comparative experiments and exhaustive ablation studies") intact. The italicised portion is the addition. The original abstract is from "Boosting Self-Supervised Graph Representation Learning via Anchor-Neighborhood Alignment and Isotropic Constraints" (avg human rating 4.33, rejected band); the abstract length grew by 40 words in the manipulated version, in line with the QUANT-ADDED condition's mean increase of 39.74 words reported in Section 5.6.

A.7. Quantitative Term List Derivation

The rate of quantitative terms in the LLM's written justifications (Table 4) is computed against a list of fifteen terms.

The corpus used for the derivation is the 7,404 ICLR 2024 abstracts, each labelled by GPT-4o-mini along two binary attributes: `quantitative_specificity` (whether the abstract mentions exact numbers, percentages, or effect sizes) and `scale_claims` (whether it mentions dataset size, model parameters, or compute requirements). The thesis defines "quantitative framing" as having either component (Section 1), so the two labels are combined with a logical OR, giving 2,193 positive labels and 5,211 negative labels.

For each content word in the corpus, we compute the log-odds ratio

$$\text{log-odds}(w) = \log \frac{P(w \mid \text{quant})}{P(w \mid \text{non-quant})}$$

where $P(w \mid \text{group})$ is the count of w in that group divided by the total content-word tokens in that group. Content words are lowercased alphabetic tokens of four or more characters, with common English function words filtered out. Words with total frequency below 200 are excluded to avoid noisy estimates. The ten words with the highest positive log-odds are retained, after filtering out LaTeX commands (such as `mathcal`, `mathrm`, `mathbb`) and corpus-specific proper nouns (such as `llama`, `imagenet`, `cifar`) that are unlikely to appear in the LLM's natural-language justifications. Five general quantitative terms that the log-odds heuristic may underweight (`accuracy`, `performance`, `dataset`, `parameter`, `scale`) are appended, with duplicates removed.

The final list of fifteen terms, in the order produced by the derivation, is:

`times`, `quantization`, `average`, `respectively`, `improvement`, `length`, `instruction`, `sota`, `pruning`, `scaling`, `accuracy`, `performance`, `dataset`, `parameter`, `scale`