QUOTE OR INVENT? FABRICATED EVIDENCE IN LLM-WRITTEN PEER REVIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimates indicate that language models now draft a visible share of peer reviews. Reviews can back their comments with quotations from the paper, i.e., words in quotation marks. The other reviewers and the authors trust a quotation in a review to provide the paper's own words. When quotations turn out to be reworded or invented, that trust is lost, and with it also the trust in the quality of peer review. We measure how often this happens. Fourteen models from Anthropic, OpenAI, Google, DeepSeek, Moonshot and Meta review 100 accepted ICLR 2026 papers under a prompt that requires a verbatim quotation for every strength and weakness. In one condition the model receives the paper. In the other, the same paper with numbered sentences, and every quotation must cite the sentence it comes from. A deterministic checker locates every quotation in the text the model read, and three LLM judges from vendors other than the reviewing model's grade every quotation the checker cannot confirm. 30 % of the 2,759 usable reviews contain at least one quotation that is not the paper's wording, but the models are far apart. With the paper alone, the rate runs from 2 and 3 % for Gemini 3.1 Pro and GPT-6 Astra to 70 % for gpt-oss-120b, and the share of reviews with a quotation whose meaning differs from the paper from 0 to 16 %. Numbering the sentences and asking for the number lowers the rate by 3.7 points. The quote-back check of a deployed review assistant would pass 73 % of the invented quotations. The fuzzy matcher of earlier work accepts every changed number, negation and changed digit.

1 INTRODUCTION

Pangram Labs reports that its classifier labeled 15,899 (21 %) of the ICLR 2026 reviews, as AI-generated (Emi, 2025). Scientific venues have started using language models in the review process: ICLR 2025 had a model comment on reviews (Thakkar et al., 2025), and AAAI-26 gave each of 22,977 papers one clearly labeled AI review (Biswas et al., 2026).

False quoted text undermines trust in reviews. Reviews provide feedback and comments which are supported with quoting from the paper. A reviewer who writes that the authors "claim X" and puts X in quotation marks tells the authors, the other reviewers and the chairs that the paper says X. If the paper does not, the review used false information to back up its comments. Every such false quotation in a review raises a doubt if the comment is correct and undermines the trust in the review process. False quotations are not limited to LLMs, they occur in human reviews too. HalluPeer (Lin et al., 2026) measured whether detectors find such false quotations. But no previous research measured how often reviewers, human or model, misquote the paper they review.

We study integrity of quotations in LLM-written reviews. In this work, we ask how often LLM-written reviews contain quotations that are not in the paper, and to what extent this differs across models. Our focus is only on what a review puts in quotation marks, we do not check every statement a review makes about the paper. Judging the paper's content, e.g., whether a review is right that a baseline is missing or that a result does not hold, is an orthogonal line of research, that studies human reviews (see Related Work, Section 2).

Checking if a quotation occurs in a paper is hard. Intuitively, the answer to the question, whether a quoted string occurs in a paper, seems deterministic: just compare the text. A plain string comparison cannot measure this, because it detects only whether, and by how many characters, a quotation differs from the text the model read. Indeed in our research we find that exact matching rejects

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

86,% of quotations that were not changed at all (Section 6). The reason is that a quotation can differ from the text it comes from. In a text extracted from the PDF, line breaks split words, ligatures become special characters and formulas come out garbled, so a correct quotation may not match, although the model quoted the paper correctly. These quotations also need judgment, which is the hard part. Because once the text comparison fails, you only know that the quotation is not identical to the paper. You do not however know why the quotation differs, if the difference matters and if the quotation exists at all or was invented. A quotation not found by the string comparison may still be correct, because the model harmlessly edited (a capital letter or punctuation) or reworded (while keeping its original meaning). But it may also be fabricated or incorrectly quoted, if the model changed the wording or the meaning. Under the rule the prompt sets, every case is an error, because the review claimed that the text was from the paper but provided a different quote. Measuring this is not trivial. Differentiating between these cases requires locating and reading the passage the quotation came from. Our research shows that even careful readers disagree on it: the judgments of our five human annotators agree, on whether a quotation departs from the paper, for only 65,% to 82,% of such quotations, and individual LLM judges call between 19,% and 49,% of the same quotations invented.

How we carry out our study. Every model receives the same instruction: for each comment provide a quotation copied word for word from the paper, and do not use double quotation marks for anything else. Everything that a review puts in double quotation marks therefore claims to be the paper's own words. We check each such claim. We run the study in two conditions. In the first, the model receives the paper as it is. In the second, every sentence of the paper carries a number, and the model must give the number of the sentence each quotation comes from. The second condition tests whether a model invents fewer quotations when it must provide, for each quotation, the number of the sentence it comes from. To give the number, the model has to find the sentence in the paper. This condition changes several things at once (identifiers in the text, a longer input and the task of citing), so it tests the package and not just a single mechanism. On edits of real quotations with known answers, the judges flag 95.8,% of the harmful edits and 2.7,% of the harmless ones (Section 4). Each of the five annotators takes the judges' side on 70,% to 87,% of the quotations the checker could not confirm, about as often as the annotators agree with one another (65,% to 82,%). The people call more of these quotations invented than the judges do (31,% to 37,% against 29,%), so, measured against any of them, our rates err low. Of the quotations the checker accepted, the people call at most 3 of about 60 invented (Section 6).

Implications of our results for using language models in reviews. *First*, invented quotations are not occasional noise but a property of the model, differing widely between models. Under the same instruction, Gemini 3.1 Pro and GPT-6 Astra put an invented quotation into 2,% and 3,% of their reviews and gpt-oss-120b into 70,%. Interestingly, requesting sentence numbers lowers the rate by only 3.7 points. This is surprising, because the expectation is that when the model is given the sentence number, if it has found it, copying it should be the easy part, so the invention rate should drop sharply. We found it fell only from about 30 % to about 26 %. The lesson: telling a model to cite its source does not make it copy. What would work is a mechanical step, in which the model gives only the sentence number, and the system inserts the sentence text itself, this is what the "Deterministic Quoting" in Related Work Section 2 does. Our data show that models find the right text passage almost always and mainly fail to copy it. Whether a model-written review can be trusted to quote the paper therefore depends mostly on which model wrote it. Using a better prompt does not improve the results. *Second*, current tools do not reveal the problem. Published benchmarks do not predict this well because they score whether claims are supported, but not whether the wording is from the paper, and a general capability index leaves a factor of three to ten between models of equal rank (Section 2). The quote-back check of a deployed review assistant would pass 73,% of the invented quotations, and fuzzy matching accepts every changed number. Relying on these tools will result in overlooking the fabricated quotations in its reviews. *Third*, our study demonstrates that this problem can be measured and checked at scale. Our checker confirms 88,% of the quotations automatically. To test it, the annotators checked them and found at most 3 of about 60 of these not to be the paper's wording. The lesson: the checker is almost always right when it says a quotation is the paper's wording, and the automatically accepted 88 % hide very few errors. Even the "at most 3" do not exclusively contain checker mistakes, e.g., some are grading noise, for example a person treating repaired extraction damage as rewording. LLM judges grade the rest and agree with people about as well as people agree with one another. *Fourth*, the judges are milder than each of the five annotators, so, measured against any of them, the rates we report err low rather than high.

Contributions. We carried out the measurement of quotation fidelity in reviews written by fourteen current models on 100 ICLR 2026 papers and 36,386 quotations. The rate of reviews with an invented quotation spans a factor of 35 between models. Lower tier models invent more and we find no association with paper length. We also developed a checker for quotations against extracted PDF text, with a taxonomy of extraction damage and editing, validated on 2,938 controlled edits. It refers 1,707 of the 1,708 harmful edits to grading and one of the 1,230 harmless ones, where exact matching rejects 86 % of unchanged quotations and fuzzy matching accepts every changed number.

Organization. We compare our research to related work in Section 2. We then describe the design in Section 3, explain the measurement methodology in Section 4, report the results in Section 5 and analyze the checker in Section 6. We discuss the limitations in Section 7 and conclude in Section 8.

2 Related work

The question we explore in this paper, falls between two lines of research, which is one reason why it has been overlooked. Work on human reviews checks whether reviewers' claims are supported by the paper (Ghorbanpour et al., 2026; Ou et al., 2025; Daghighfarsoodeh et al., 2026), but it does not check quotations. Work on quotations checks the claims by exact or near string match, as one component of a larger system and on clean text: quotations from a story in AI debate (Khan et al., 2024), sentences copied from a retrieval context (Long et al., 2026), and Deterministic Quoting replaces the model's quotation with a stored passage (Yeung, 2024). GopherCite constrained decoding to verbatim quotes from retrieved documents (Menick et al., 2022). They does not look at reviews, and its checks were not built for PDF format. The question has also only recently become pressing. A misquotation used to be one reviewer's error in one review. Now a detector labels a fifth of the reviews at a major venue as fully AI-generated (Emi, 2025), and venues increasingly put models into the review process (Thakkar et al., 2025; Biswas et al., 2026). A model that tends to misquote does so across every review it drafts, so how often each model misquotes determines the extent at which a venue's reviews can be trusted.

Hallucinations in reviews and in papers. HalluPeer injects hallucinations into human-written ICLR and NeurIPS reviews and benchmarks detectors on them. Using two annotators (Lin et al., 2026) found 20 genuine hallucinations in 1,161 human NeurIPS 2024 reviews. ClaimCheck grounds reviewers' weaknesses to the claims a paper makes (Ou et al., 2025), and the concurrent Peerify benchmarks the support of 800 reviewer claims by the manuscript (Daghighfarsoodeh et al., 2026). Both study human reviews and semantic claims. We study LLM reviews and quotation fidelity, which, unlike a claim, can be checked against the text in paper. Its accuracy can therefore be measured automatically at scale and compared across models. Checking a claim needs expert judgment, which is slow and uncertain: The best verifier in Peerispect (Ghorbanpour et al., 2026) reaches an accuracy of only 0.287 on human reviewer claims. For references, an automatic audit of 48,095 accepted papers found at least two likely hallucinated reference in 2.5 % of ICLR 2025 papers (Russinovich et al., 2026).

LLMs as judges. Without a reference, judges agree with experts mainly on items they could answer themselves, and expert-written references largely remove that dependence (Krumdick et al., 2026). Our judges always see the relevant passages of the paper, they do not grade the reviews of their own vendor and are tested against known answers before use.

What published benchmarks predict. Grounding benchmarks ask whether a response's claims are supported by a document the model was given. FACTS Grounding marks an answer accurate only if every claim in it is grounded in that context (Jacovi et al., 2025), and its current version keeps the data and updates the judges (Cheng et al., 2025). That criterion does not score wording: a paraphrase inside quotation marks is grounded in the sense FACTS measures. Nor does the leaderboard (Google DeepMind and Kaggle, 2026) cover the field. Of our fourteen models it lists four (GPT-6 Astra 82.0 %, Gemini 3.8 Flash 73.1 %, Gemini 3.1 Pro 65.7 %, gpt-oss-120b 30.7 %, read 22 September 2026) and no model of the Claude 5 generation. A general capability index reaches thirteen of the fourteen. Against the Artificial Analysis Intelligence Index (Artificial Analysis, 2026) we find our rate per quotation correlating at Spearman -0.72 ($p = 0.007$). Still, the index is a poor guide for choosing a model. Models it ranks alike differ three- to tenfold: GPT-6 Astra and Claude Fable 5.1 have the same index value, yet Fable invents quotations more than three times as often, and Gemini 3.1 Pro ranks below Claude Sonnet 5 but invents a tenth as often. The index predicts the share of

reviews with an invented quotation even more weakly (-0.51, $p = 0.08$), because that share also depends on how often a model quotes, which no benchmark reports. It runs from 4.7 quotations per review for Llama 4 Maverick to 29.9 for Claude Opus 5. From the existing leaderboards, a reader could guess roughly how the models rank, but not how often they invent a quotation.

What our research adds. The deployments at ICLR 2025 and AAAI-26 each ran one review system, built on one vendor's models, as reviewer's assistant or reviewer (Thakkar et al., 2025; Biswas et al., 2026). We compare fourteen models from six vendors, frontier and open-weight, as reviewers under one prompt, and check each of their 36,386 quotations individually. Fourteen models from six providers produced 2,800 reviews under different output limits, refusal behavior and version changes, and excluding the reviewer's vendor from the jury required judges from four vendors. Of them 2,759 were used. Earlier quotation checks were not validated against known answers (Khan et al., 2024; Long et al., 2026; Menick et al., 2022). The reason is that quotation grading has no existing ground truth. We therefore built two: real quotations we changed on purpose, so that the correct grade is known, and five people who each graded the same 300 quotations. In contrast, ours is validated on 2,938 controlled edits, and on text extracted from PDFs rather than clean text. Earlier LLM judges also grade the outputs of their own vendors and correct for self-preference afterwards (Jacovi et al., 2025; Krumdick et al., 2026), our judges never do. Studies that check a model against people use one to three annotators per item and merge their labels (Ou et al., 2025; Lin et al., 2026; Daghighfarsoodeh et al., 2026; Krumdick et al., 2026). As a result, they cannot tell how much of the error belongs to the model and how much to the annotator. In our research five people graded all 300 of our validation items independently and we report each separately. This allows us to find out how much people disagree and separates the judges' errors from the annotators'.

3 STUDY DESIGN

This section describes the design of our study. To rule out an analysis tuned to the results, we fixed the design, hypotheses and analysis in a protocol before collecting any data.

Papers. We use 5,346 accepted ICLR 2026 papers in our corpus. We exclude papers whose extracted text exceeds 40,000 words, so that every prompt fits the smallest context window in the roster, and exclude papers with fewer than 2,000 words before the references. Our resulting corpus contains 5,309 papers. They fall into four strata at the quartiles of total length, appendices included. We drew 100 papers at random, 25 from each length quartile (6,894 to 27,816 words, median 13,464), so that short and long papers are equally represented. We convert every PDF with `pdftotext` in reading order and give the model that text.

Models. Fourteen models review each paper (Table 4): four Claude models (Fable 5.1, Opus 5, Sonnet 5, Haiku 4.5), GPT-6 Astra and GPT-5.6 Luna, Gemini 3.1 Pro and Gemini 3.8 Flash, and six open-weight models served on Microsoft Foundry (DeepSeek-V4-Pro and -Flash, Kimi K2.6, gpt-oss-120b, Llama 4 Maverick, Llama 3.3 70B). We run the models with their defaults. We send no temperature and no reasoning setting, because the study asks how deployed models behave.

Prompt and conditions. One prompt asks for an ICLR-style review and sets one rule for evidence. Every strength and weakness is supported by at least one quotation of the paper, copied word for word in double quotation marks, with omissions marked by an ellipsis. We do not use double quotation marks for anything else. In C1 the model receives the paper. In C2 every sentence carries an identifier such as `[[S12]]`, and each quotation must cite the sentence it comes from. Removing the identifiers restores the C1 text character for character.Each model is scheduled to review each paper once per condition, 2,800 reviews; for 20 papers each model writes two more C1 reviews to measure variation between calls.

Outcomes and hypotheses. A quotation of six or more words is *invented* if its wording departs from the paper beyond cosmetic differences, that is, grade 2 or higher on the scale of Section 4. The primary outcome Y of a completed review is 1 if at least one of its quotations is invented; Y_3 uses the stricter cut at a changed meaning or a fabrication. H1: in C1, $P(Y = 1)$ differs between the fourteen models. H2: $P(Y = 1)$ differs between C2 and C1. Both are tested two-sided with paper-level permutation tests and Holm's correction (Section 5).

4 MEASURING QUOTATION FIDELITY

Finding quotations. A quotation is a span of two or more words between double quotation marks, paired within a paragraph or list item. Curly marks keep their direction, so a quotation inside a quotation stays inside it. Spans under six words are recorded but kept out of every outcome, because a short phrase can match the paper or fail by accident. Khan et al. (2024) draw a similar line at eight words when they count fake quotes. Our heuristic shows that using six words is more inclusive, while avoiding accidental matches.

Automatic classes. First, before comparing, our checker cleans up both texts and normalizes the quotation and the paper the same way: it evens out things that differ for harmless reasons, like, ligatures, typographic quotes and dashes, invisible characters, line-break hyphens, running headers and page numbers. This step does not touch the numbers because numbers are where a small edit changes the result the most, but the judges would not detect this. Therefore a quotation with a changed number is not accepted. It then assigns each span a class (Table 5). Two classes are accepted automatically: *exact*, a quotation that occurs word for word in the paper, and *elided*, a quotation that marks omissions with an ellipsis and whose pieces all occur in the paper, in order and close together. We run a substring match, so a quotation may begin or end inside a word: the quotation "stable estimator in the main algorithm" would be accepted as exact in a paper that says "an unstable estimator in the main algorithm", although it reverses the meaning. To see whether this happens, we located 27,423 of the 30,609 exact spans again in the text and checked their word boundaries. We found that 23 begin or end inside a word, all where PDF extraction joins a table or figure to the running text, and they do not reverse a meaning. All other classes are routed to the judges: spans that match only after digits are dropped, or only with single quotes read as double or Greek letters spelled out, fragments out of order, pieces separated by a figure caption, and spans not in the paper at all (*altered* if, for the worst of its fragments, the longest shared run covers at least 40 % of that fragment, *fabricated* otherwise).

What our checker adds. Earlier quotation checks (see Section 2) answer one question with yes or no, and they control how the model quotes: in tags (Khan et al., 2024), as whole copied sentences (Long et al., 2026), or under decoding constrained to verbatim text (Menick et al., 2022). But in practice reviews do not have a fixed format nor clean text, they contain free text that quotes extracted PDF text, and a yes/no check cannot differentiate a repaired line break from a changed number, or a reworded quotation from an invented one. On extracted text it fails in both directions (Table 2). We built our checker on the quote-back gate of a deployed review assistant and, while developing it, found five ways in which that gate accepts changed text, among them a normalizer that deletes digits, so that "IPv4" matches "IPv6" (Appendix C). Our checker finds quotations in free text by pairing double quotation marks, accepts omissions marked with an ellipsis, and repairs the damage of PDF extraction without changing a digit. What it cannot accept it sorts into classes, among them a match that holds only once digits are dropped, fragments out of order, and text with no counterpart in the paper, and the judges then grade each routed span from verbatim to fabricated.

Grading. We grade every routed span: 0 verbatim; 1 cosmetic (case, punctuation, a dropped citation marker, a repaired extraction artifact, a marked editorial change); 2 reworded with the meaning kept; 3 meaning changed (a number, a sign, a negation, a hedge, a scope or a comparison); 4 fabricated; U undecidable. Three judges grade each span, drawn from Claude Opus 5, GPT-6 Astra, Gemini 3.1 Pro and DeepSeek-V4-Pro, with the restriction that they do not grade its own vendor's reviews. A judge sees the span, the words before it in the review, and up to 3 passages of the paper around the checker's closest match and the span's rarest words. It does not see the model, the condition or the rest of the review. A reply counts only if at least half of the 3-word sequences of the words it names as matched, and one run of 4 words, occur in the passages, which lets a judge repair extraction damage while tying its evidence to the passages. The span's grade is the median of the three. If one judge fails or answers U, the lower of the other two counts, and two U make the span undecidable.

Validation. Before the main run, the judges graded synthetic edits of real exact quotations with known answers: 8 harmful edits (a changed number, a negation, a replaced or deleted content word, a sentence joined without an ellipsis, a sentence from another paper, swapped fragments, a changed digit next to letters) and 5 harmless ones by construction (case, typographic marks, a ligature, a hyphenated line break, a marked omission). The labels are format edits and a marked omission can change a meaning, as one of the 300 does by dropping a "but not for". The preregistered test

required a sensitivity and a specificity of at least 90 % at the cut at 2 and a sensitivity of at least 80 % at the cut at 3. The judges reached 95.8 % (543 of 567), 97.3 % (400 of 411) and 98.9 % (173 of 175), and each of the 4 trios cleared all thresholds, so they graded the main run. The trios are not equally sensitive. The two that grade the OpenAI and the Gemini models catch 91 % of the harmful edits, the two that grade the other houses 98 and 99 %. After the main run, 5 members of the project graded the same 300 blinded spans independently: 240 routed spans drawn equally from the 4 bands of the judges' grade and 60 spans the checker accepted. We compare each with the judges separately.

Why three judges. As individuals the judges disagree far more than their median suggests. On the 609 C1 spans that Claude Opus 5, Gemini 3.1 Pro and DeepSeek-V4-Pro all graded numerically, Gemini calls 48.9 % invented, Opus 29.6 % and DeepSeek 18.7 %; on the 714 C1 spans of the trio that has GPT-6 Astra in place of DeepSeek, Astra and Gemini call 41.0 and 40.5 % invented against Opus's 26.5 % (an addition to the preregistered analysis, as is the rest of this paragraph). The constructed answers put the same models in the same order: GPT-6 Astra and Gemini 3.1 Pro grade 99 % of the harmful edits as invented (428 of 431 and 496 of 498 graded numerically), Claude Opus 5 85 % and DeepSeek-V4-Pro 40 %. The preregistered test required not each judge, but each trio of judges to pass, and every trio with DeepSeek-V4-Pro passed, because the other two outvote it. Apart from DeepSeek-V4-Pro, every judge grades a changed number, a negation or a sentence from another paper as invented in 87 to 100 % of the cases, and what separates them is whether a quotation that departs from the extracted text is a repaired artifact or a rewording. The median absorbs it. Every trio reaches a sensitivity of 91 to 99 %, and redrawing each span from the error matrix its own trio showed on the synthetic edits leaves the order of Table 1 unchanged and moves no model by more than two points. With three judges the rate becomes a property of the reviews rather than of the judge we happened to pick.

5 Evaluation Results

Of the 2,800 scheduled cells, 2,759 produced a usable review; 34 stopped at the output limit even after the re-run rule (all of them Llama 4 Maverick) and 7 were refusals. The reviews contain 36,386 quotations of six or more words; 20 further spans with an unpaired quotation mark were recorded. The checker confirmed 32,003 of them and sent 4,383 to the judges, which graded 325 verbatim, 2,754 cosmetic, 1,078 reworded, 67 changed in meaning and 122 fabricated; 30 could not be decided and 7 did not receive two usable grades.

How often (RQ1). 30.2 % of the usable reviews contain at least one invented quotation. The models differ far more than that average suggests (Table 1). In C1 they run from 2 % (Gemini 3.1 Pro) and 3 % (GPT-6 Astra, Gemini 3.8 Flash) to 70 % (gpt-oss-120b), 61 % (Claude Haiku 4.5) and 56 % (DeepSeek-V4-Flash). The permutation test rejects equal risk across the fourteen models ($T_1 = 70.5$, $p = 0.0001$, Holm 0.0002). Per quotation the spread is wider, from 0.3 % (GPT-6 Astra, Gemini 3.1 Pro) to 19.6 % (Llama 3.3 70B). The severity variant keeps the same models at the clean end and reorders the other end: in C1, $Y_3 = 0$ for GPT-6 Astra and Gemini 3.1 Pro against 16 % for Claude Haiku 4.5, 14 % for Llama 3.3 70B and 10 % for gpt-oss-120b ($p = 0.0001$).

Numbered sentences (RQ2). Asking for the sentence number lowers the rate by 3.7 points ($[-6.6, -1.0]$, $p = 0.012$, Holm 0.012), and by 2.0 points for Y_3 ($[-3.5, -0.5]$, $p = 0.010$). The effect is not uniform: GPT-5.6 Luna falls from 42 to 23 % and DeepSeek-V4-Flash from 56 to 43 %, while the two cleanest models reach zero, Claude Haiku 4.5 rises slightly, and Llama 4 Maverick's Y_3 rises from 1 to 9 %. The cited identifier covers the quotation's place for 99 % of GPT-6 Astra's faithful quotations (verbatim or cosmetic), 98 % of Gemini 3.1 Pro's, but only 47 % of gpt-oss-120b's and 66 % of Llama 4 Maverick's.

Length and tiers (RQ3). We find no association between paper length and the rate. The slope over the four length strata is $+0.7$ points per stratum ($p = 0.51$, $[-1.6, +3.0]$), and the stratum means lie between 29.5 and 35.2 %. Within a vendor, the stronger model invents less. GPT-6 Astra is 39 points below GPT-5.6 Luna ($p < 0.001$); the Claude ladder falls by 16 points per step from Haiku 4.5 to Fable 5.1 ($p < 0.001$); DeepSeek-V4-Pro is 16 points below DeepSeek-V4-Flash ($p = 0.032$). The two Gemini models do not differ (1 point, $p = 1.0$), so the pattern is not a law.

What the deployed gate would catch (S5). We replayed every invented quotation through the quote-back component of the deployed gate to see what a production check misses. It would pass

Table 1: Reviews with at least one invented quotation (Y), with a changed meaning or a fabrication (Y_3), quotations of six or more words per review, the share of them the judges grade invented, and the share of C2 quotations graded faithful whose cited sentence number covers the quotation (S4). Rates in per cent; quotations per review and the rate of invented quotations per quotation are for C1; C1 is the paper alone, C2 the paper with numbered sentences.

	Y (%)		Y_3 (%)		quotations, C1		S4
Model	C1	C2	C1	C2	per review	invented (%)	correct id
Claude Fable 5.1	17	19	9	5	22.8	0.9	91
Claude Opus 5	16	18	9	8	29.9	0.8	91
Claude Sonnet 5	41	40	11	7	18.0	2.9	80
Claude Haiku 4.5	61	66	16	12	5.9	17.9	76
GPT-6 Astra	3	0	0	0	11.9	0.3	99
GPT-5.6 Luna	42	23	2	4	18.1	3.1	88
Gemini 3.1 Pro	2	0	0	0	6.6	0.3	98
Gemini 3.8 Flash	3	3	1	1	9.1	0.3	96
DeepSeek-V4-Pro	40	39	5	1	8.9	5.7	85
DeepSeek-V4-Flash	56	43	8	5	8.6	9.2	74
Kimi K2.6	13	9	1	0	9.2	1.6	89
gpt-oss-120b	70	61	10	2	11.2	12.7	47
Llama 4 Maverick	28	25	1	9	4.7	7.7	66
Llama 3.3 70B	55	50	14	8	5.7	19.6	69

73 % of them, and in 41 % of the affected reviews it raises no misquotation at all. For the quotations whose meaning changed or that were fabricated, it passes 86 %.

Variation between calls. Three calls of the same model on the same paper give the same Y for the clean models (5 % of the repeat papers differ for GPT-6 Astra and Gemini 3.1 Pro) and far less often for the rest, on 10 % to 80 % of the repeat papers; 80 % differ for Claude Haiku 4.5, DeepSeek-V4-Flash and Llama 3.3 70B. Llama 4 Maverick completed all three calls on only 7 of the 20 papers and Claude Fable 5.1 on 19. We therefore draw no conclusion about a model from one review, and report every rate over all completed reviews.

6 Validating the checker (RQ4)

Table 2 shows what the checker and the simpler checks do with 2,938 controlled edits of real quotations. Our checker sends 1,707 of the 1,708 harmful edits to grading to the judges: every changed number, negation, changed digit inside a token, replaced or deleted content word, later sentence joined on and sentence from another paper, and 299 of 300 swapped fragments. Of the 1,230 harmless edits it sends only one, which is a correctly marked omission. It accepts every lowered first letter, typographic mark, ligature and hyphenated line break. Of the edits that change what a quotation says (a changed number, a negation, and so on), the judges gave 95 % to 100 %, depending on the type of edit, a grade of 2 or higher. They recognized them as not the paper's wording. This is the judges' sensitivity at the cut at 2, per edit type (Table 2, "judges" column). The one exception is a later sentence joined on without an ellipsis, which they call invented in 84 % of the cases.

The simpler checks fail in both directions. Exact matching on the extracted text accepts only 14 % of *unchanged* quotations, because a line break inside the quotation defeats it, and collapsing whitespace lifts that to 78 %. The string match of Khan et al. (2024) accepts 96 % of unchanged quotations but rejects nearly every harmless repair of extraction damage: 5 % for typographic marks, 3 % for a ligature, none for a hyphenated line break or a marked omission. Fuzzy matching at the threshold the Copy-Paste paper states (Long et al., 2026) accepts all unchanged quotations and, with them, 100 % of changed numbers, 98 % of negations and changed digits, 61 % of deleted and 50 % of replaced content words. The threshold its code uses, 0.7, accepts most harmful edits: 100 % of changed numbers, negations, changed digits and replaced or deleted content words, 80 % of a later sentence joined on, 54 % of swapped fragments, and one of 300 sentences from another paper. Table 2 tests only whether that code's matcher accepts a candidate. The Copy-Paste pipeline then

Table 2: Synthetic edits of real quotations (per cent). "Referred" is the share the checker sends to grading, "judges" the share the three judges grade as intended (invented for the harmful edits, faithful for the harmless ones), the remaining columns the share each simpler check accepts as verified: exact match on the extracted text, the same with whitespace collapsed (+ws), the string match of Khan et al. (2024), and Copy-Paste (Long et al., 2026) at the threshold its paper states (CP .95) and as the code does it (CP code). The first eight edits are harmful, the last six harmless.

Edit	referred	judges	exact	+ws	Khan	CP .95	CP code
changed number	100	95	0	0	0	100	100
negation	100	100	0	0	0	98	100
digit in a token	100	100	0	0	0	98	100
sentence from another paper	100	100	0	0	0	0	0
content word replaced	100	99	0	0	0	50	100
content word deleted	100	96	0	0	0	61	100
later sentence joined on	100	84	0	0	0	3	80
fragments swapped	100	97	0	0	0	0	54
unchanged	0	99	14	78	96	100	100
lowered first letter	0	98	0	4	96	100	100
typographic marks	0	97	0	4	5	100	100
ligature	0	100	2	3	3	99	100
hyphenated line break	0	99	0	0	0	99	100
marked omission	0	93	0	0	0	4	94

replaces an accepted candidate with the matched source text, which we do not evaluate. A reviewer who checks quotations this way accepts the edits that change what the paper is said to claim.

Against people. Table 3 compares the judges with each of the five annotators A1-A5. The analysis was run once per annotator. The people disagree with one another. Pair by pair they take the same side of the cut at 2 on 65 % to 82 % of the routed spans (weighted back to the run), and each takes the judges' side on 70 %, 87 %, 82 %, 75 % and 80 %. Agreement with the judges is thus of the same order as agreement between people, and agreement alone does not show who is right. Disagreements run both ways, but on balance the people are stricter. They call 36 %, 37 %, 35 %, 34 % and 31 % of the routed spans invented, the judges 29 %. At the cut at 3 the weighted sensitivity is only 9 % to 31 %, but it is based on few spans. 1 to 13 quotations the judges graded verbatim or cosmetic, and some they graded as rewordings, each stand for many spans of their band. Counted the judges put 67 % to 87 % of the quotations a person reads as a changed meaning at 3 or 4. The lowest sensitivity at the cut at 2, 48 %, comes from the annotator who moved the most spans the judges had graded verbatim or cosmetic above the cut, 16 of 60 against 8, 10, 13 and 9. Read against the extracted text at least 12 of those 16 repair extraction damage or drop a citation marker, which the scale grades as cosmetic.

The annotators also locate the one failure of the judges that the protocol anticipated: when the passage search misses, a real quotation is graded as fabricated. This happens mostly with mathematical notation, which PDF extraction garbles. A quarter of the sampled spans the judges graded fabricated are mostly formulas (15 of 60, by a fixed count of mathematical symbols, and the annotators identified at most four of them fabricated. The error inflates the count of fabricated quotations, the rarest class, and, where a formula was quoted correctly, the invented rate Y. On balance the judges nonetheless remain milder than the annotators. Therefore our recommendation is if a quotation consists mostly of mathematical notation, do not trust the LLM judges' verdict on it, have a person check it instead. The audit supports the checker's automatic acceptance. Of the 60 quotations the checker accepted, the five annotators together flag six as invented, each by one or two annotators, and read against the extracted text, all six are verbatim or correctly elided. The last row of Table 3 carries these flags into the rates as given. The simpler checks fail on real quotations as they do on the synthetic edits: exact matching rejects 78 % to 79 % of the quotations the annotators call faithful, and Copy-Paste at the threshold of its code accepts 80 % to 92 % of those they call invented.

7 Discussion

Interpreting quotation fidelity. Quotation marks in a review make a claim that the quoted text corresponds to the text in the paper. Y counts the reviews in which at least one such claim is

Table 3: The judges against each of five annotators (per cent). Sensitivity and specificity carry 95 % intervals from 2,000 bootstrap resamples within bands; all shares are weighted back to the run by the population of each band of the judges' grade. The last four rows are the registered sensitivity analyses with that annotator as reference, 1,000 replicates each: "judge error" redraws every graded routed span from the annotator's error rates, independently per span (Dirichlet posterior with Jeffreys prior); "checker error" flips accepted spans at the audit's rate (Beta posterior with Jeffreys prior). They give the mean over the fourteen models of the posterior-median C1 rate (registered 32 %) and H2 in points with 95 % intervals (registered -3.7 for Y, -2.0 for Y_3).

	A1	A2	A3	A4	A5
sensitivity, cut 2	48 [38, 62]	73 [61, 88]	66 [55, 80]	54 [43, 69]	65 [53, 82]
specificity, cut 2	82 [77, 87]	95 [92, 97]	91 [87, 94]	85 [81, 89]	87 [83, 91]
sensitivity, cut 3	22 [12, 54]	17 [11, 25]	31 [19, 55]	9 [7, 14]	22 [13, 43]
specificity, cut 3	97 [97, 97]	98 [98, 99]	98 [97, 98]	98 [98, 99]	98 [97, 98]
share invented (person)	36	37	35	34	31
same side of cut 2 as judges	70	87	82	75	80
audit: invented / accepted	1 / 59	3 / 60	0 / 60	2 / 59	2 / 60
mean C1 rate, judge error	40	39	39	39	36
H2 (Y), judge error	-3.2	-3.6	-3.4	-3.3	-3.4
95 % interval	[-5.9, -0.5]	[-5.5, -1.5]	[-5.8, -1.3]	[-5.9, -0.7]	[-5.7, -0.8]
H2 (Y_3), judge error	-1.3	-2.5	-1.6	-2.7	-1.7
95 % interval	[-3.5, +0.9]	[-4.8, -0.1]	[-3.6, +0.2]	[-5.6, +0.1]	[-4.1, +0.5]
mean C1 rate, checker error	45	60	35	54	53

false. What a false claim costs depends on what it changes. Most of the quotations Y counts, keep the meaning of the text, i.e., they misrepresent the wording, but not the content. Nevertheless this undermines the trust: a reader cannot tell a faithful quotation from a reworded one without checking, but exactly that work a quotation was meant to save. The quotations that change the meaning or do not exist in the paper (Y3) misrepresent the content as well. In those cases the review's evidence does not correspond to the paper's. Because a reworded quotation can leave the content intact, a review can be right while misquoting, and a verbatim quotation can support a wrong criticism.

Measurement uncertainty. Our rates carry the errors of the checker, the passage search and the judges. In general the judges are milder than the annotators. Because one wrongly accepted quotation flips a whole review, the checker error rate, that an audit of 60 acceptances cannot rule out, raises the mean C1 rate by up to 28 points.

8 Conclusion

Quotations help readers assess whether a source supports a reviewer's comments. But that benefit holds as long as the quoted passage accurately reproduces the source. If a model reconstructs the text passage, its comments and its purported evidence can contain the same error and appear to support one another. A changed qualification, number, or comparison can make the manuscript appear to support a claim it does not support. We evaluated quotation fidelity in reviews generated by fourteen language models for 100 scientific manuscripts. The models put at least one non-verbatim quotation into 30 % of their reviews, from 2 % to 70 % depending on the model. They almost always find the right passage (99.6 % of quotations have a counterpart in the paper) but they fail at copying it, mostly without changing the meaning. For applications of models for reviews, the model plays a far more central role than the prompt. Our research shows that quotations can be checked at scale: our checker settles 88 % automatically and leaves the rest to judges that agree with people about as often as people agree with one another. The conclusions that can be derived from our research indicate that for research on reviews, checking quotations complements checking claims: whether a quoted string occurs in the paper has a definite answer and can be measured across many models, while whether a claim is right needs expert judgment. For research that uses language models as judges, one human annotator does not suffice to validate them. As our research shows people disagree with one another about as often as with the judges.

REFERENCES

Artificial Analysis. Artificial analysis intelligence index. https://artificialanalysis.ai/leaderboards/models, 2026. Read 22 September 2026.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. AI-assisted peer review at scale: The AAAI-26 AI review pilot, 2026. Preprint, v1 of 15 April 2026.

Aileen Cheng, Alon Jacovi, Amir Globerson, Ben Golan, Charles Kwong, Chris Alberti, Connie Tao, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Yonatan Bitton, Sasha Goldshtein, and Dipanjan Das. The FACTS leaderboard: A comprehensive benchmark for large language model factuality, 2025. Preprint, v1 of 11 December 2025; 66 authors, first eleven and last two listed.

Alireza Daghighfarsoodeh, Sajad Ebrahimi, Ali Ghorbanpour, Soroush Sadeghian, Radin Cheraghi, Negar Arabzadeh, and Ebrahim Bagheri. Peerify: Benchmarking peer-review claim verification, 2026. Preprint, v1 of 2 September 2026.

Bradley Emi. Pangram predicts 21% of ICLR reviews are AI-generated. Pangram Labs. https://www.pangram.com/blog/pangram-predicts-21-of-iclr-reviews-are-ai-generated, accessed 2026-09-12, 11 2025. Vendor analysis.

Ali Ghorbanpour, Soroush Sadeghian, Alireza Daghighfarsoodeh, Sajad Ebrahimi, Negar Arabzadeh, Seyed Mohammad Hosseini, and Ebrahim Bagheri. Peerispect: Claim verification in scientific peer reviews. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, 2026. doi: 10.1145/3805712.3808368.

Google DeepMind and Kaggle. FACTS grounding leaderboard. https://www.kaggle.com/benchmarks/google/facts-grounding/leaderboard, 2026. Leaderboard dated 11 September 2026, read 22 and 23 September 2026.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The FACTS grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input, 2025. Preprint, v1 of 6 January 2025.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23662–23733, 2024.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. No free labels: Limitations of LLM-as-a-judge without human grounding. In *Conference on Language Modeling (COLM 2026)*, 2026.

Tzu-Ling Lin, Dong-Ting Yao, Teng-Fang Hsiao, Wei-Chih Chen, and Hong-Han Shuai. HalluPeer: A taxonomy-driven benchmark for detecting hallucinations in scientific peer reviews, 2026. Preprint, v1 of 3 September 2026; accepted to Findings of EMNLP 2026.

Yongchao Long, Yingying Zhang, Xianbin Wen, Xian Wu, Yuxi Zhou, and Shenda Hong. Copy-paste to mitigate large language model hallucinations. In *International Conference on Learning Representations (ICLR 2026)*, 2026.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes, 2022. Preprint, v1 of 21 March 2022.

Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. ClaimCheck: How grounded are LLM critiques of scientific papers? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21712–21735, 2025. doi:10.18653/v1/2025.findings-emnlp.1185.

Mark Russinovich, Ram Shankar Siva Kumar, and Ahmed Salem. Phantom references: Hallucinated citations that survive peer review at top-tier conferences, 2026. Preprint, v2 of 6 July 2026.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025, 2025. Preprint, v1 of 13 April 2025; journal version: Nature Machine Intelligence 8(3):326–336, 2026, doi:10.1038/s42256-026-01188-x.

Matt Yeung. Deterministic quoting: Making LLMs safer for healthcare. `https://mattyyeung.github.io/deterministic-quoting`, accessed 2026-09-18, 4 2024.

A AI USE STATEMENT

In this research we study language models and use them for research. We describe their use in Section 3 and Section 4. Beyond that, we used an AI coding assistant for tasks, and it provided feedback on the research methodology. We also used language models for tasks that include improving our writing in this paper. We used it for reviewing the protocol and the code against the primary sources in our corpus. The authors checked every finding of that review against the sources before accepting it, verified every reference against its primary source, and tested the code with unit tests and integration runs. We take responsibility for the final content of this work.

B ETHICS STATEMENT

Our research uses papers published in the ICLR 2026 proceedings. No unpublished or confidential manuscript was used in the study. The generated reviews were not sent to authors, to ICLR or to any venue. The annotators are members of the project and of the research group.

C THE CHECKER AND THE DEPLOYED GATE

When developing the checker we found five ways in which text is accepted which the paper does not contain, and reproduced on real reviews:

1. its normalizer deletes digits, so “IPv4” matches “IPv6” and a changed number in a line of results passes;
2. its check for an interrupted quotation accepts a deletion inside the quotation, including the deletion of a negation;
3. it ignores the order of the fragments between marked omissions, and one-word fragments entirely;
4. a quotation attributed to the paper but bearing little resemblance to it is reported, not failed;
5. an elision matches under the digit-deleting normalization, so adding an ellipsis to a quotation with a changed digit makes it pass.

The study’s checker requires the strict normalization that keeps every digit on both the whole span and every fragment, it keeps the order of fragments, and it routes every class except *exact* and *elided* to grading. The classes are listed in Table 5.

Table 4: The fourteen models. "ID" is the identifier sent to the provider; the output limit counts reasoning and visible text together.

Model	Provider	ID	Output limit
Claude Fable 5.1	Anthropic	`claude-fable-5-1`	32,000
Claude Opus 5	Anthropic	`claude-opus-5`	32,000
Claude Sonnet 5	Anthropic	`claude-sonnet-5`	32,000
Claude Haiku 4.5	Anthropic	`claude-haiku-4-5`	16,000
GPT-6 Astra	OpenAI	`gpt-6-astra`	32,000
GPT-5.6 Luna	OpenAI	`gpt-5.6-luna`	32,000
Gemini 3.1 Pro	Google	`gemini-3.1-pro-preview`	32,000
Gemini 3.8 Flash	Google	`gemini-3.8-flash`	32,000
DeepSeek-V4-Pro	Foundry	`DeepSeek-V4-Pro`	32,000
DeepSeek-V4-Flash	Foundry	`DeepSeek-V4-Flash`	32,000
Kimi K2.6	Foundry	`Kimi-K2.6`	32,000
gpt-oss-120b	Foundry	`gpt-oss-120b`	32,000
Llama 4 Maverick	Foundry	`Llama-4-Maverick-17B-128E-Instruct-FP8`	8,192
Llama 3.3 70B	Foundry	`Llama-3.3-70B-Instruct`	8,192

Table 5: Automatic classes of a quoted span. "Strict" is the normalization that keeps every digit except page numbers; "gate" is the deployed gate's normalization, which can delete digits. Every class except *exact* and *elided* is routed to the judges.

Class	Rule	Routed
exact	occurs in the paper under the gate's and the strict normalization	no
exact-normalized	occurs only under the gate's normalization	yes
exact-lenient	occurs once single quotes are read as double, Greek letters spelled out, [s] and [es] dropped	yes
elided	every fragment between marked omissions occurs, in order, each within 1,500 characters of the previous, under both normalizations	no
elided-normalized	as elided, but only under the gate's normalization	yes
elided-unordered	every fragment occurs, but not in order or not that close	yes
interrupted	the paper has the span in intact pieces with other text between them	yes
interleaved	the span's parts occur in order near its best match, separated in the paper by whitespace or at least 15 characters of other text	yes
altered	not in the paper; the longest shared run covers at least 40 % of the worst fragment	yes
fabricated	not in the paper; that run covers less than 40 %	yes

D MODELS AND CLASSES

E REGISTERED SECONDARY RESULTS

Table 6 shows: the 95 % paper-bootstrap interval of every C1 rate, the per-model difference C3 − C1 with its interval, the share of C1 reviews that quote at all, the C3 rate per quotation, and the share of C3 identifiers that point to a neighboring sentence (S4 adjacent). Three registered sensitivity analyses leave H2 where it is: counting every ungraded span as invented gives -3.8 points, taking the upper of two grades where one judge failed -3.5, and scoring the 41 unusable cells as 0 or as 1 gives -3.6 and -3.7. For Y_3 the length slope is $+1.1$ points per stratum ($p = 0.12$). The pilot on 20 development papers identified 94 of 274 completed reviews as invented. In the repeated analysis, Llama 4 Maverick has all three calls on 7 of the 20 papers and Claude Fable 5.1 on 19; the others have all 20.

Table 6: Registered secondary results per model. Y C1 with its 95 % paper-bootstrap interval; C3 − C1 in points with its interval; the share of C1 reviews with at least one quotation of six or more words; the rate of invented quotations per quotation in C3; and the share of C3 quotations graded faithful whose cited identifier names the sentence next to the quotation's place (S4 adjacent).

Model	Y C1	95 %	C3 − C1	95 %	quoting, C1	per quot., C3	S4 adj.
Claude Fable 5.1	17	[10, 25]	2	[−8, 11]	100	0.9	0
Claude Opus 5	16	[9, 23]	2	[−8, 12]	100	0.6	0
Claude Sonnet 5	41	[31, 51]	−1	[−14, 14]	100	2.4	0
Claude Haiku 4.5	61	[52, 70]	5	[−7, 16]	96	12.9	2
GPT-6 Astra	3	[0, 7]	−3	[−7, 0]	100	0.0	0
GPT-5.6 Luna	42	[32, 52]	−19	[−33, −5]	100	1.4	0
Gemini 3.1 Pro	2	[0, 5]	−2	[−5, 0]	100	0.0	0
Gemini 3.8 Flash	3	[0, 7]	0	[−4, 4]	100	0.3	1
DeepSeek-V4-Pro	40	[30, 50]	−1	[−13, 12]	100	5.7	1
DeepSeek-V4-Flash	56	[46, 66]	−13	[−26, 0]	100	5.9	2
Kimi K2.6	13	[7, 20]	−4	[−12, 4]	96	0.9	2
gpt-oss-120b	70	[61, 79]	−9	[−21, 4]	100	10.4	3
Llama 4 Maverick	28	[19, 39]	−6	[−22, 9]	89	6.3	5
Llama 3.3 70B	55	[45, 65]	−5	[−18, 8]	84	14.0	15