Language Models Should Be Used to Surface the Unwritten Code of Science

Honglin Bao[†], Siyang Wu[*], Jiwoong Choi[*], Yingrong Mao[*], James A. Evans

University of Chicago

* SW, JC and YM contributed equally.

† Correspondence: honglinbao@uchicago.edu

Abstract

This position paper calls on the research community not only to investigate how human biases are inherited by large language models (LLMs) but also to explore how these biases in LLMs can be leveraged to make society's "unwritten code" — such as implicit stereotypes and heuristics — visible and accessible for critique. We introduce a conceptual framework through a case study in science: uncovering hidden rules in peer review - the factors that reviewers care about but rarely state explicitly due to normative scientific expectations. The idea of the framework is to push LLMs to speak out their heuristics through generating self-consistent hypotheses - why one paper appeared stronger in reviewer scoring - among paired papers submitted to 46 academic conferences, while iteratively searching deeper hypotheses from remaining pairs where existing hypotheses cannot explain. We observed that LLMs' normative priors about the internal characteristics of good science extracted from their self-talk, e.g., theoretical rigor, were systematically updated toward posteriors that emphasize storytelling about external connections, such as how the work is positioned and connected within and across literatures. Human reviewers tend to explicitly reward aspects that moderately align with LLMs' normative priors (correlation = 0.49) but avoid articulating contextualization and storytelling posteriors in their review comments (correlation = –0.14), despite giving implicit reward to them with positive scores. These patterns are robust across different models and out-of-sample judgments. We discuss the broad applicability of our proposed framework, leveraging LLMs as diagnostic tools to amplify and surface the tacit codes underlying human society, enabling public discussion of revealed values and more precisely targeted responsible AI.

1 Introduction

Large language models (LLMs), extensively trained on vast corpora of human-generated text and visuals, inevitably absorb and reflect the patterns of human thought embedded within these datasets. Researchers have explored how AI can be used to judge human activities such as scientific research (Liang et al., 2024), misinformation oversight (Chen et al., 2024a), and perceptions of physical appearance (Conwell et al., 2025). With strong predictive capacities, LLMs can reflect public opinion and human beliefs on contentious political issues, including same-sex marriage (Kim and Lee, 2023) and climate change (Lee et al., 2024b). These models often reveal biases that closely align with directly measured human biases and even amplify them across various social dimensions (Ziems et al., 2024). This position paper moves beyond merely acknowledging LLM biases, and it calls on leveraging them as diagnostic tools to uncover and examine the implicit, unwritten evaluative codes that govern evaluative judgments in science.

The "unwritten code" – comprising tacit norms, implicit stereotypes, and subtle conventions that invisibly influence societal dynamics – represents a potent force that, if left unexamined, unsummarized, and unaddressed, can perpetuate systemic inequities and injustices. We use the term narrowly: domain-level evaluative heuristics that predict decisions but are absent from stated criteria and articulated feedback. They may be one channel through which broader social structures, e.g., linguistic or institutional privilege (Bourdieu, 1988; Hariharan, 2019), operate; we identify the heuristics, not their structural causes. Leveraging the inherent biases embedded within LLMs can shed light on otherwise elusive pat-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2264–2289

November 6-10, 2026 ©2026 Association for Computational Linguistics

terns that sustain societal inequality. One straightforward path to mitigating human bias involves summarizing and clearly articulating the biases humans hold, subsequently enabling evaluation, regulation and systemic change. Nevertheless, these biases frequently exist "unwritten" and thus resist formalization and systematic analysis. Even though some methods can expose bias in LLMs when models appear to avoid expressing it, as by analyzing associated words (Bai et al., 2025), there remains a deeper layer of unwritten code that contemporary simple techniques cannot recover. These deeper structures activate during the model's reasoning process, serving as heuristics that streamline its efficiency [1]. An algorithmic framework capable of extracting and explicitly articulating them through self-reinforcement and amplification of the rationale within LLMs, which are inherently imbued with human biases and heuristics, could illuminate the content and character of bias within models and human cultures. Thus, this position paper calls for an intentional shift in perspective: **We call on researchers not only to investigate how human biases are inherited by LLMs but also to explore how these biases can be leveraged to make the unwritten codes in science visible**, as shown in Figure 1. Peer review is a suitable first test: it has a documented gap between stated criteria and judgment (Teplitskiy, 2016), large-scale scored decisions, and textual artifacts; other settings remain a research agenda. To this end, we present a conceptual framework that enables this process by pushing LLMs to speak out their heuristics through searching their deeper self-consistent hypotheses that can explain their decision-making. By prompting models to think out of the box and "exaggerate" hidden rules that humans might be unable or unwilling to acknowledge openly, we can more clearly illustrate how such biases and heuristics operate and which groups they potentially disadvantage.

2 Related Work

Learned representations have long been used to study human culture and bias (Caliskan et al., 2017; Garg et al., 2018; Kozlowski et al., 2019; Blodgett et al., 2020), including through indirect probes (Bai et al., 2025; Chen et al., 2025; Fang et al., 2024). These works mainly measure static associations, i.e., a model's *prior*. We instead study the prior-to-posterior shift produced by hypothesis search over scored decisions (Zhou et al., 2024; Liu et al., 2025), using LLMs as diagnostic tools rather than substitute reviewers (Liang et al., 2024). See Appendix G for more.

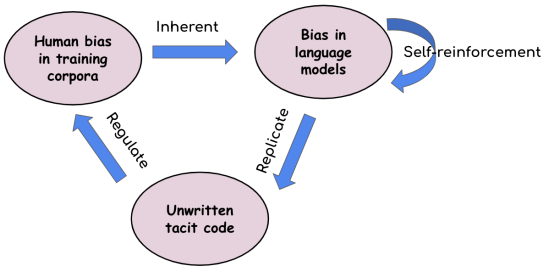


Figure 1: We argue and present a conceptual framework to self-reinforce biases to amplify unwritten codes until the model articulates and replicates them explicitly.

3 A Case Study: The Aesthetic of Peer Review

What makes a scientific paper "good"? While scientific communities often emphasize norms such as rigor, the actual process of evaluating quality is far more entangled with social dynamics, implicit biases, and tacit expectations (Bao and Teplitskiy, 2024). Peer review, the central gatekeeping mechanism of science, is deeply shaped by these social dynamics. Despite its appearance of objectivity, peer review can be biased toward well-known institutions, established authors, or fashionable topics (Lee et al., 2013). Reviewers may assign high scores to work that aligns with unspoken disciplinary tastes or collective values (Sulik et al., 2025), while "lazily" masking these preferences in the language of standardized evaluation criteria (Purkayastha et al., 2025). As a result, much of what constitutes the tacit code of a field – its evaluative culture – remains hidden beneath the surface, shaping scientific recognition in nontransparent ways that remain hard to alter.

Data: We collected data from the OpenReview API and Paper Copilot [2], a website that aggregates

[1] We have more discussions in Section 6, Section 2, and Appendix G.

[2] https://papercopilot.com/

peer review scores (mostly) for computer science conference submissions. Our final dataset includes metadata, peer review scores, and review comments for 26,731 papers submitted to 46 conferences including computer science, physics, medicine, and social sciences. We use peer review scores as a proxy for the paper's perceived quality, as shown in Appendix A, due to the consistency and high confidence of judgments among reviewers.

To represent each paper's content, we follow the approach of (Zhang et al., 2025), constructing an extended abstract that integrates additional structured elements — contextual background, key ideas, methodological and theoretical details, experiments and results, and mentioned impact. We generate these extended abstracts using prompts used in (Zhang et al., 2025) with GPT-4o. This representation generation has been empirically shown to achieve high quality across evaluation metrics and to closely match human annotations. It also demonstrates strong alignment with the original text and a low rate of hallucination (Amar et al., 2023; Zhang et al., 2025), see Appendix E for fidelity experiments. We deliberately avoid feeding full texts directly to LLMs for two main reasons. First, context windows remain limited, preventing broad coverage that would enable more generalizable hypotheses. Current approaches, e.g., (Movva et al., 2025), employ reduction methods such as sparse autoencoders to compress long inputs into activated neurons for hypothesis generation, but these are typically applied only to straightforward cases (e.g., political attitudes) where activations can be reduced to obvious dimensions (e.g., gender and race). Second, prior work has shown that LLMs often struggle with long, unstructured inputs, leading to degraded performance and increased hallucination rates mid-text (Liu et al., 2024b; Yen et al., 2024). When presented with lengthy texts, attention may diffuse across irrelevant sections, impairing reasoning over core content. To address these issues, we adopt a curated and condensed input format, ensuring that the model attends to the most relevant information. Dataset and paper representations are detailed in Appendix H. Most importantly, robustness checks in the Appendix E show that LLM judgments on full papers and extended abstracts yield highly consistent results (consistency = 0.89).

4 The Framework: Prior, Posterior, and Hypothesis Generation

We adopt a hypothesis generation framework – a topic increasingly explored in the automation of scientific discovery using AI (Radensky et al., 2024; Liu et al., 2024a; Batista and Ross, 2024; Liu et al., 2025) – to update LLMs' stereotyped priors about good science into posteriors that reflect how human evaluators and LLMs actually judge. Priors are hypotheses generated via self-talk without data; posteriors are hypotheses found by the iterative search below, weighted by the share of pairs each explains.

Hypothesis Generation: We begin by constructing a pairwise dataset. For each conference, we identify all pairs of papers where one paper has received a significantly higher review score than the other. To reduce noise due to randomness, we define a "significant" score gap as one that exceeds one standard deviation within the score distribution of that conference. From this pool, we randomly sample 50 paper pairs (due to the limited contextual window) and feed them jointly into an LLM. The LLM is tasked with reading, understanding, and comparing the papers' extended abstracts, and then generating five hypotheses to explain why the better scoring paper may appear stronger than the other. Joint input encourages generalizable hypotheses.

Self-consistency Refinement: Each new hypothesis is then evaluated one pair at a time on all 5,000 pairs. For each pair and hypothesis, we ask the LLM to determine the degree to which the hypothesis explains the observed quality difference (human review scores) using the following prompt: "Analyze the content from both papers to determine whether Paper 2 is superior in the aspects identified in the hypothesis (i.e., if the comparison of content supports the hypothesis)." This type of self-confirmation explicitly reveals the inherent biases of LLMs during the judgment process. The LLM outputs two things: (1) a binary label (0 or 1) indicating whether the hypothesis explains the score difference, and (2) a confidence score (0–10) for its judgment. We use 3-fold voting and compute the final label using a confidence-weighted scheme: if the sum of confidence scores for the label "1" exceeds that of label "0", the final label is 1; otherwise, it is 0. This improves self-consistency (Taubenfeld et al., 2025); majority vot-

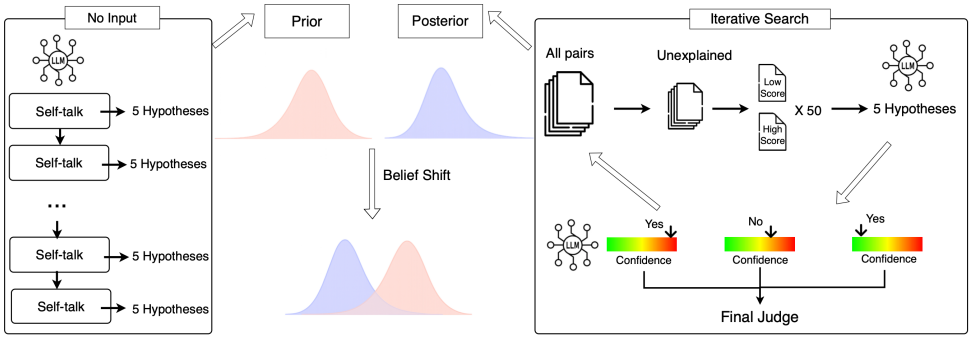


Figure 2: Updating priors to posteriors by hypothesis search. Final judge: confidence-weighted voting over three randomized-order judgments per pair and hypothesis.

ing gives near-identical labels (r > 0.98). To limit position bias (Shi et al., 2024), we use large score gaps and randomize paper order (77% consistency; Appendix M). A hypothesis's *posterior coverage* is the share of pairs it explains (Algorithm 1).

Iterative Search: The initial set of hypotheses may not account for all observed differences. For cases where none of the hypotheses yield a confident explanation (i.e., the confidence-weighted label is 0 for all hypotheses), we designate them as "unexplained". In each round, we sample 50 such unexplained pairs and prompt the LLM to generate five new hypotheses that are distinct from the existing ones. This iterative process continues until the proportion of unexplained cases drops below 5%. Note that each newly generated hypothesis is evaluated across all paper pairs, not just the subset of unexplained instances from which it was derived. Although new hypotheses are generated from a specific subset, they may still provide valid explanations for other pairs previously explained by different hypotheses. In our experiments, we find that 4 rounds of search, yielding 20 hypotheses, cover 97% of pairs under judge model agreements. Results are stable across models (89% agreement) and samples (r = 0.80; Appendices C, D). This suggests that scientific judgments can often be explained by a compact set of generalizable hypotheses, reflecting recurring evaluative patterns across diverse cases [3].

Prior and Posterior: From a Bayesian perspective, querying an LLM with a general question "What do you think makes one paper appear stronger than another?" without providing any specific data input, elicits the model's **prior** $P(\text{hypothesis})$. This prior reflects the LLM's internal distribution over explanatory hypotheses about what makes a scientific paper strong, independent of evidence. To extract this prior, we prompted the model to generate hypotheses in the absence of data. Specifically, we conducted 4 rounds per simulation, with 5 distinct hypotheses generated in each round, as in the posterior search. Repeating this simulation 250 times yielded a total of 5,000 prior hypotheses, following the same procedure of iteratively searching hypotheses. When the model is shown actual data, such as comparisons between paper pairs, it updates its beliefs, resulting in a **posterior** distribution $P(\text{hypothesis} \mid \text{data})$ over hypotheses, which captures how likely the model believes a given hypothesis explains the observed data and how confidently it makes the judgment. Operationally, we estimate this posterior by computing, for each hypothesis, the proportion of paper pairs it can successfully explain when the model is exposed to actual inputs.

We are particularly interested in belief change during this process, i.e., how LLM's prior about good science transforms into a robust posterior, after judging the actual pairwise dataset. For each generated hypothesis, when the LLM judges the pairwise dataset, we retain the proportion of cases it can explain (one pair can be explained by multiple hypotheses). We also match each hypothesis to priors (prompting: do you think they are talking about the same idea and same hypothesis?) to obtain their appearance frequency in the prior. We also conducted

[3] In Section 6 and the results below, we further decompose these generalized hypotheses into priors (which humans tend to explicitly emphasize) and posteriors (humans' implicit judgments).

cross-validation experiments for the reliability of LLM labels of "match" using a heuristic dictionary-based method (locating some keywords of posterior hypotheses in priors to identify "match"), and these two methods exhibit a relatively high degree of correlation of 0.70, p=0.001 (see Appendix K for details). The frequency of each posterior hypothesis in the prior distribution is measured by counting how often a similar hypothesis appears within each window of 20 sequential prior hypotheses across 250 simulations. We use 20-hypothesis windows to align with the posterior generation setup, where LLMs are prompted to produce 4 rounds of 5 unique hypotheses. This ensures a fair comparison between the frequency of hypotheses in the prior and posterior distributions.

Overall experiments are formalized in Appendix F. Across experiments, we used OpenAI o3-mini for hypotheses (naively generate priors and iteratively generate posteriors in hypothesis search). We used OpenAI GPT-4o to match prior and posterior hypotheses, i.e., using a good reasoning model to "think" and a high-capacity aligned model to give a stable and reliable judgment. The overall framework of this process can be seen in Figure 2. All prompts used in the paper can be viewed in Appendix I. Generating high-quality summaries for roughly 27K papers, pairing them under our criteria to create more than 14 million pairs, and having an advanced LLM to read, hypothesize, and evaluate each pair multiple rounds would lead to prohibitive computational and monetary costs. For demonstration purposes in our case study, we randomly selected 5,000 pairs for the main experiments. To strengthen the robustness of our conclusions, we further evaluated an additional 5,000 pairs through two robustness checks: (1) whether they produce similar hypothesis patterns, and (2) whether the generated set of hypotheses based on one dataset can be applied to out-of-sample data. Detailed discussions in Appendix D show that our conclusions are both generalizable and robust. Furthermore, in Appendix C we replicate our experiments across a range of popular closed- and open-source LLMs, all of which yield qualitatively similar patterns.

Early (prior-like): "Compared to the other, one paper ... not theoretically rigorous."
Late (posterior): "Compared to the other, one paper ... not interact well with adjacent domains."
Full list: Table 1.

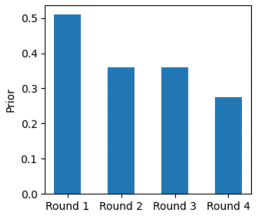
5 Results

We highlight the key results in the main paper. To comprehensively test generalizability and robustness, we designed 13 sections of experiments in the appendix. Across models, out-of-sample data, human annotation, and alternative representations of papers, we find that the major conclusions consistently hold.

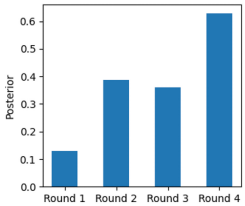
5.1 Prior and posterior manifest distinct generation logics

We first examine the 20 posterior hypotheses by search round (Figure 3). As shown in Figure 3a, the frequency with which hypotheses appear in the model's prior belief gradually decreases across rounds (represented by the mean of each round). This trend makes intuitive sense: the LLM initially generates hypotheses that are most "normative" for the standard of good science, such as theoretical rigor. These are the kinds of stereotyped considerations that naturally come to mind when researchers judge scientific work. As the rounds progress, we prompt the model to think beyond common ideas, encouraging the generation of less frequently seen hypotheses that can still explain peer review decisions in previously unexplained cases. Interestingly, these later hypotheses exhibit stronger explanatory power across the dataset. That is, although they are generated from cases that were initially hard to explain, they nevertheless prove applicable to cases already covered by earlier, more normative hypotheses. Moreover, the LLM shows greater confidence — quantified as the difference between the total confidence of votes matching the final weighted label and that of votes for the alternative label across three rounds — and improved consistency, measured by the proportion of pairs in which LLM casts the same vote across all rounds. Initially, the judging LLM draws on its priors but lacks confidence in assessing factors such

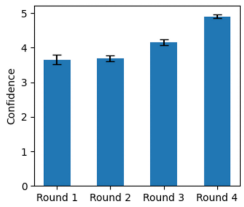
as theoretical rigor. The model shows greater confidence in later rounds when evaluating the newly generated hypotheses, however, which are largely centered on presentation, contextualization, and storytelling (we will return to these hypotheses in the next section). The full list of generated hypotheses is presented in Table 1 in the appendix. These hypotheses are also distinct, with the pairwise cosine similarity in posterior coverage (1–0 sequences) being 0.43 ± 0.02, indicating that they cover relatively different dimensions of the scientific quality signal.



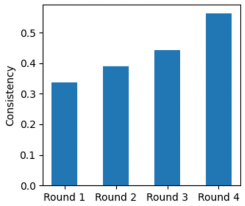
(a) The appearance of the generated hypothesis in the priors



(b) The proportion of explained pairs by generated hypothesis (posterior) across rounds



(c) LLM's increased confidence when making the judgment



(d) LLM's increased consistency when making the judgment

Figure 3: Prior and posterior manifest distinct generation logics. Means per search round over the 20 posterior hypotheses; (a) prior frequency via matching to 250 prior simulations, (b) posterior coverage.

5.2 Human judgments and LLMs share priors, but not posteriors

We observe notable shifts in the frequency of certain hypotheses from prior to posterior, as shown in Figure 4. Some hypotheses remain relatively stable in frequency and their explanatory power for good science (middle). In contrast, others manifest substantial change – some decrease markedly in frequency (left), while others increase significantly (right) from the prior distribution (see Table 1). This shift reflects a transition, after being exposed to the pairwise data, from emphasizing normative scientific values – such as theoretical or methodological rigor - to prioritizing more practical, communicative, and contextualization aspects, including storytelling quality and interdisciplinary relevance. When treated as a peer reviewer, the LLM initially aligns with conventional scientific ideals. In practice, however, it leans more heavily on accessible, narrative-driven criteria that appear more explanatory in distinguishing lower from higher rated papers (Figure 3 panels c and d). This suggests a divergence between explicit and implicit criteria of evaluation in science, a phenomenon widely appreciated although rarely measured in social studies of science (Latour, 1987).

We collected human review comments and annotated them using in-context learning prompts with GPT-4o, leveraging its strong performance in social computing annotation tasks, often surpassing human experts and crowd-workers (Törnberg, 2023; Alizadeh et al., 2023; Haq et al., 2025). For each hypothesis, the annotation model was provided with three illustrative examples of reviewers not-mentioning (label 0) specific aspects in a hypothesis, along with the reviewers' attitudes toward those aspects (positive: 1 or negative: -1, they are considered as "mentioning"). For robustness, we conducted two experiments: 1) the dictionary method produces high consistency with LLM's "mention" label with correlation of 0.78, $p = 0.000$, detailed in Appendix K. 2) Three master's students specializing in social computing independently conducted the human annotations, achieving a high level of consistency with the LLM-generated labels (see Appendix B for details). We find human comments align with LLM's priors in terms of the frequency of aspects they address, meaning that they articulate a similar attention distribution as LLM's prior. In contrast, human reviewers rarely articulate their updated judgments like the LLM does. That is, if we assume the LLM mirrors human reviewing behavior, both begin from similar priors. However, while the LLM transparently updates its evaluation (posterior), human reviewers often refrain from articulating such shifts – likely due to normative expectations in academic reviewing – despite favoring unarticulated values with higher scores (we have more discussions and empirical evidence in Section 6).

Supporting this, the top five hypotheses in Table 1 with the largest prior-to-posterior gains (e.g., story-

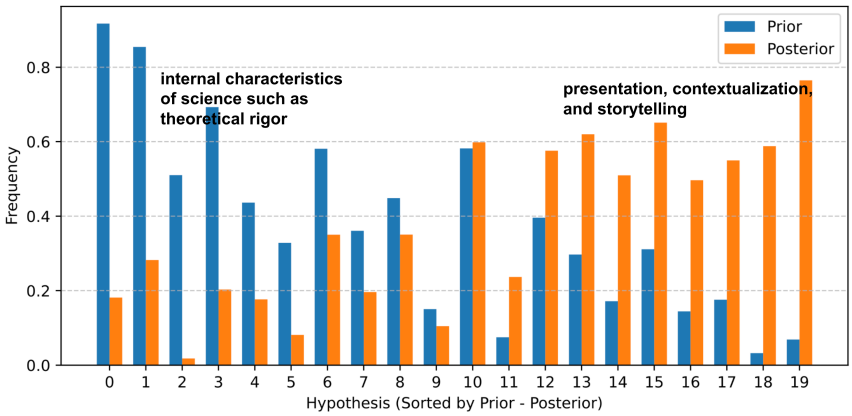


Figure 4: Prior versus Posterior. The index of each hypothesis represents its rank in the change of the prior frequency minus the posterior frequency (from high to low).

telling, orange texts) account for 61% of the LLM's posterior attention, but only 15% of its prior and 21% of mentions in human feedback. Conversely, the top five hypotheses with the largest prior-to-posterior losses (e.g., internal rigor, blue texts) account for 17% of the LLM's posteriors, but as much as 68% of its prior and 50% of human mentions. LLMs, as expected, initially even exaggerate stereotyped evaluation norms — overweighting "prior" and underweighting "posterior" standards — more sharply than humans do. Humans indeed explicitly reward internal qualities, but their stated reward patterns explain very little of the actual score distribution (R^2 = 0.06, details about the regression analysis can be seen in Appendix L), suggesting the presence of implicit rules guiding the evaluation process. Human reviewers tend to explicitly comment on aspects that moderately align with the LLM's normative priors (mention frequency and prior frequency correlation = 0.49) but are much less likely to articulate storytelling and contextualization posteriors in their written feedback (mention frequency and posterior coverage correlation = –0.14). All these suggest a possibility that while these storytelling aspects gain importance during evaluation — filling in the missing explanatory power R^2 of regressions — they remain underreported in human reviews, perhaps because they deviate from traditional scholarly norms, despite their relevance for establishing significance.

6 Discussions

It is important to emphasize that our goal is not to advocate for the automation of judgment tasks by LLMs. Rather, our primary goal is to **present a position through a conceptual framework** — namely, to exploit the biases and preference rationale embedded in LLMs by human discursive culture, then self-amplifying, reproducing, and explicitly identifying who might be harmed by these biases. Much of the existing research has focused on more explicit, easily identifiable forms of biases associated with social dimensions such as gender and race (Ntoutsi et al., 2020). Researchers have developed methods to bypass the barriers when the model refuses to disclose its biases, such as using associated words (Bai et al., 2025). However, they constitute the "prior" itself. These biases and heuristics still manifest superficially and can be uncovered through direct or indirect cues. In contrast, what we focus on in this paper is the "posterior" and how it diverges from the prior. Surprisingly, the results show that even when an LLM's prior (i.e., its stereotypes) is "good", its posterior can still be "bad" - let alone when its prior is already "bad". When acting as reviewers, LLMs appear to draw on two distinct systems to explain differences in quality signals. One is more normative and less confident in the evaluation process. The other is more narrative-driven, context-sensitive, and ultimately more confident and consistent. When judgment can be achieved through both systems, LLMs are more likely to apply the second

system across the board as heuristics because they more consistently explain quality differentials. Humans, like LLMs, may also favor contextual criteria in evaluation, but they articulate this shift in terms of normative standards.

We make our position dialectical with alternative perspectives and policy suggestions:

Is narrative quality a bias or a prerequisite? Reviewers cannot judge the rigor of a paper they cannot understand, so narrative quality is partly endogenous to evaluation. It plays a dual role: an epistemic prerequisite and a potential source of inequity, as rhetorical conventions track linguistic and institutional privilege (Chen et al., 2024b). It is "hidden" because scores reflect it while reviews rarely mention it, and several posteriors (contextualization, adjacent domains, background knowledge) go beyond readability. Separating clarity from storytelling preference requires causal tests we leave to future work.

How to articulate the unwritten code in other high-stakes evaluative settings? These are possible extensions, not demonstrated generalizations. In hiring, LLMs can be asked to rank resumes: while initial judgments may rely on normative criteria such as education, experience, and skills, prompting models to articulate deeper rationales can surface less explicit but influential factors like perceived cultural fit—for instance, inferring misalignment when a candidate's prior employers differ culturally from the hiring firm (Rivera, 2012) and being rejected silently. Similarly, in college admissions, encouraging LLMs to explain their evaluations of personal statements enables systematic analysis of implicit reasoning; prior work shows that essays from certain ethnic or linguistic backgrounds exhibit stylistic differences that may be inadvertently penalized (Lee et al., 2025; Alvero et al., 2021; Alvero and Pattichis, 2024). Examining LLM scoring rationales in these contexts can thus potentially help identify how implicit signals contribute to bias in hiring, admissions, and criminal justice.

To what extent does an LLM's shift from prior to posterior actually reflect human reasoning? Humans may never explicitly confess to using narrative or contextual heuristics in reviewing and other social dimensions, yet convergent but indirect supporting evidence suggests that they do. Prior survey-based studies show that reviewers in the social sciences place substantially greater emphasis on "interpretive challenges that question a study's framing" than on methodological rigor (Teplitskiy, 2016; Strang and Siler, 2015). We push LLMs to articulate such emphasis and generalize this insight to conference papers in technical fields, where - owing to normative expectations around methodological rigor and SOTA performance - reviewers are less inclined to make such concerns explicit [4] but use "lazy" words as a mask precisely in NLP (Purkayastha et al., 2025). Moreover, if human score variance could be understood as comprising both explicit and implicit components, with LLMs' prior-to-posterior shift accounting for nearly all variance (as shown in Section 4 "iterative search") while human-stated explicit judgments map only to the prior, then the posterior can reasonably be taken to capture the implicit part. This evidence is indirect: no ground truth confirms that the extracted hypotheses are reviewers' actual criteria, so our framework surfaces *candidate* codes; direct validation (e.g., controlled experiments, reviewer think-alouds) is future work.

Another view holds that even if we acknowledge the existence of unwritten codes, **making them explicit serves little purpose and may even enable people to "game the system."** But this view neglects the crucial balance between transparency and normativity. The fear that publicizing unwritten codes will be exploited implicitly assumes that it is preferable for such rules to continue operating in the shadows. In reality, the opposite is true: "hidden curriculum" benefits those who already occupy advantaged positions or happen to be in the know (Hariharan, 2019). For example, researchers from elite institutions who are familiar with Western academic norms naturally understand how to write and frame a compelling "story" (Chen et al., 2024b) - they have long benefited from tacit rules without needing them to be made explicit. The purpose of transparency is not to provide a cheat sheet, but to allow marginalized groups to understand the hidden barriers they face and thereby push institutions to reflect on and revise those standards. If people do game the system after unwritten codes are revealed, it is not a failure of transparency — it is evidence that those codes themselves need institutional correction.

[4] https://hackingsemantics.xyz/2020/reviewing-models/

Does surfacing implicit criteria imply automating judgment? No; that question is separate from ours. Nevertheless, our work suggests that insofar as LLMs function as cultural technologies that encode societal norms (Blodgett et al., 2020; Farrell et al., 2025), then their "cognitive shift" in peer review reflects analogous unacknowledged shifts in human reviewers. That said, while LLM-generated reviews have been proven useful and cover broader and more specific aspects than human reviews (Liang et al., 2024; Yuan et al., 2022; Thakkar et al., 2025) – likely including those humans may be unwilling or unable to disclose – LLM reviewers still lack constructive feedback essential for critique when judging a standalone paper. Pairing papers for comparative judgment of feature existence and magnitude in some sense addresses this concern, as demonstrated here and in previous experiments (Bai et al., 2025).

What should we do at the implementation level? LLMs cannot be made fully unbiased, but they can be designed to surface, label, and contrast the implicit heuristics they rely on. At an implementation level, this means shifting from single-shot judgments to procedures that explicitly elicit multiple rationales, track prior–posterior shifts, and expose which criteria become dominant only after iterative comparison. Practically, LLMs should be used as diagnostic tools rather than automated judges: systems can require models to generate and evaluate multiple competing hypotheses, retain confidence signals, and highlight systematic gaps between what is explicitly articulated and what is implicitly rewarded. These signals should be logged and presented to human decision-makers, not collapsed into final scores. We should establish evaluation metrics for transparency, not just accuracy. Implementation should emphasize perspective diversity. Using multiple models, prompts, or reviewer roles (e.g., complementary–neutral–proximal perspectives) helps prevent any single normative frame from dominating. Diversity here should be enforced not only at the level of prompts, but at the representational level - through steering, architectural tweaks, or training differences that lead to meaningfully different internal representations and behaviors. Such collective calibration mechanisms allow institutions to recognize, debate, and potentially revise their unwritten codes.

Limitations

In this paper, we focus on conference papers. Future research could examine more selective, longer, and detailed review comments from journals, although such data are often subject to legal constraints. Conducting an anonymous survey across both technology and social science fields would further strengthen the proposed framework. Our study employs well-normalized and venue-controlled scores; future work could incorporate more nuanced statistical analyses on factors such as institutional prestige. Our evidence comes from a single Western, English-language academic ecosystem, and there is no direct ground truth for the extracted codes. Values not visible in text, e.g., hostility to ideas for political or cultural reasons, or author prestige, may be missed or misattributed to textual features.

Ethics Statement

This work complies with the Code of Ethics. We used AI assistants to expedite the coding process. All code snippets produced by AI assistants were verified by the first author before they were incorporated. For writing, we only used AI assistants to check grammar. No external ethics review was required, as the experiments solely analyzed scientific papers, thereby posing no potential harm to individuals across any demographic or geographical group. Our study uses LLMs to surface implicit biases and unwritten evaluative codes in peer review. While this promotes transparency and fairness, it may also risk reinforcing stereotypes if misused. We mitigate this by anonymizing data, framing the method as diagnostic rather than prescriptive, and highlighting the importance of responsible use. All experiments were conducted using two NVIDIA 40-GB A100 GPUs.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*, 101.

AJ Alvero, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L Stevens, and Benjamin W Domingue. 2021. Essay content and style are strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. *Science Advances*, 7(42):eabi9031.

AJ Alvero and Rebecca Pattichis. 2024. Multilingualism and mismatching: Spanish language usage in college admissions essays. *Poetics*, 105:101903.

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. Openasp: A benchmark for multi-document open aspect-based summarization. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Ariful Azad and Afeefa Banu. 2024. Publication trends in artificial intelligence conferences: The rise of super prolific authors. *arXiv preprint arXiv:2412.07793*.

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, pages 752–759.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.

Honglin Bao and Misha Teplitskiy. 2024. A simulation-based analysis of the impact of rhetorical citations in science. *Nature Communications*, 15(1):431.

Rafael M Batista and James Ross. 2024. Words that work: Using language to generate hypotheses. *Available at SSRN 4926398*.

Adib Bazgir, Yuwen Zhang, and 1 others. 2025. Agentichypothesis: A survey on hypothesis generation using llm systems. *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Pierre Bourdieu. 1988. *Homo Academicus*. Stanford University Press.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chen and 1 others. 2025. To labor is not to suffer: Exploration of polarity association bias in LLMs for sentiment analysis. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Hong Chen, Christopher I Rider, David Jurgens, and Misha Teplitskiy. 2024b. Geographical disparities in navigating rejection in science drive disparities in its file drawer. *Available at SSRN 4872023*.

Colin Conwell, Daniel Graham, Chelsea Boccagno, and Edward A Vessel. 2025. The perceptual primacy of feeling: Affectless visual machines explain a majority of variance in human visually evoked affect. *Proceedings of the National Academy of Sciences*, 122(4):e2306025121.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Fang and 1 others. 2024. Born differently makes a difference: Counterfactual study of bias in biography generation from a data-to-text perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. 2025. Large ai models are cultural and social technologies. *Science*, 387(6739):1153–1156.

Richard Frankel, Jared Jennings, and Joshua Lee. 2022. Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science*, 68(7):5514–5532.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan Nadler. 2024. Online images amplify gender bias. *Nature*, 626(8001):1049–1055.

Lei Guo, Chris J Vargo, Zixuan Pan, Weicong Ding, and Prakash Ishwar. 2016. Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, 93(2):332–359.

Muhammad Uzair Ul Haq, Davide Rigoni, and Alessandro Sperduti. 2025. Llms as data annotators: How close are we to human performance. *arXiv preprint arXiv:2504.15022*.

Janani Hariharan. 2019. Uncovering the hidden curriculum. *Science*, 364(6441):702–702.

Boris V Janssen, Geert Kazemier, and Marc G Besselink. 2023. The use of chatgpt and other large language models in surgical science. *BJS Open*, 7(2):zrad032.

Junsol Kim and Byungkyu Lee. 2023. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. 2025. Scientific hypothesis generation and validation: Methods, datasets, and future directions. *arXiv preprint arXiv:2505.04651*.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language generation models can cause harm: So what can we do about it? an actionable survey. *arXiv preprint arXiv:2210.07700*.

Bruno Latour. 1987. *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

Jinsook Lee, AJ Alvero, Thorsten Joachims, and Rene Kizilcec. 2025. Poor alignment and steerability of large language models: Evidence from college admission essays. *arXiv preprint arXiv:2503.20062*.

Messi HJ Lee, Jacob M Montgomery, and Calvin K Lai. 2024a. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1321–1340.

Sanguk Lee, Tai-Quan Peng, Matthew H Goldberg, Seth A Rosenthal, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz. 2024b. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024b. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108.

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

Xinru Lin and Luyang Li. 2025. Implicit bias in llms: A survey. *arXiv preprint arXiv:2503.02776*.

Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. 2025. Hypobench: Towards systematic and principled benchmarking for hypothesis generation. *arXiv preprint arXiv:2504.11524*.

Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. 2024a. Literature meets data: A synergistic approach to hypothesis generation. *arXiv preprint arXiv:2410.17309*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Jens Ludwig and Sendhil Mullainathan. 2024. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

Benjamin S. Manning, Kehang Zhu, and John J. Horton. 2024. Automated social science: Language models as scientist and subjects. *National Bureau of Economic Research Working Paper*, (w32381).

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. Sparse autoencoders for hypothesis generation. *ICML*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, and 1 others. 2020. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356.

Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2025. Lazyreview: A dataset for uncovering lazy thinking in nlp peer reviews. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*.

Lauren A Rivera. 2012. Hiring as cultural matching: The case of elite professional service firms. *American Sociological Review*, 77(6):999–1022.

Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program chairs' report on peer review at acl 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

David Strang and Kyle Siler. 2015. Revising as reframing: Original submissions versus published papers in administrative science quarterly, 2005 to 2009. *Sociological Theory*, 33(1):71–96.

Justin Sulik, Nakwon Rim, Elizabeth Pontikes, James Evans, and Gary Lupyan. 2025. Differences in psychologists' cognitive traits are associated with scientific divides. *Nature Human Behaviour*, pages 1–15.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

Misha Teplitskiy. 2016. Frame search and re-search: How quantitative sociological articles change during peer review. *The American Sociologist*, 47(2):264–288.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, and 1 others. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. Learning to generate novel scientific directions with contextualized literature-based discovery. *arXiv preprint arXiv:2305.14259*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Christopher Wolfram and Aaron Schein. 2025. Layers at similar depths generate similar activations across llm architectures. *COLM*.

Siyang Wu, Honglin Bao, Sida Li, Ari Holtzman, and James A Evans. 2025. Mapping overlaps in benchmarks through perplexity in the wild. *arXiv preprint arXiv:2509.23488*.

Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. 2024. Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv preprint arXiv:2411.02382*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. *Findings of the Association for Computational Linguistics: ACL*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

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

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and 1 others. 2025. Massw: A new dataset and benchmark tasks for ai-assisted scientific workflows. *Association for Computational Linguistics: NAACL*.

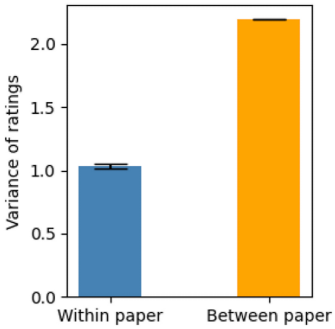
Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

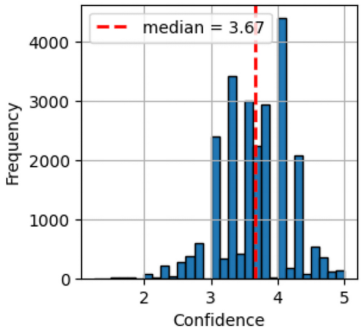
Appendix

A Reviewer score is a good measure of quality signal

Review scores tend to be much more consistent within a single paper than across different papers (see Figure 5 panel a), suggesting that multiple reviewers held shared values, enabling consensus in their evaluations. Reviewers reported relatively high confidence in their judgments with a median of 3.67 out of 5 (see Figure 5 panel b). Thus, in general the review score is a good measure of quality signal.



(a) The judgment is consistent



(b) The judgment is confident

Figure 5: Review score is a good measure for quality.

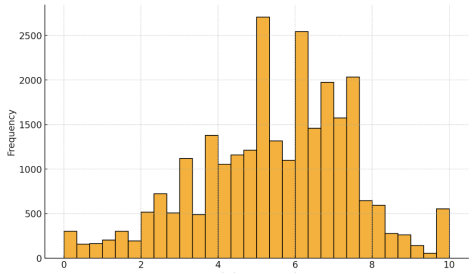


Figure 6: Review scores are normally distributed.

B Human annotation experiments

Here, we complement the labeling methods (LLM labeling (major) and the dictionary method (complementary)) with human annotation. Three master's students specializing in social computing were involved to label the reviewer comments using exactly the same prompt as GPT-4o (if reviewers' attitudes toward the aspects in a given hypothesis were positive = 1 or negative = –1, they were considered as "mentioning", or 0 - not mentioned at all). The annotators are from Asia: two males and one female. They labeled independently, without using AI assistance and without communicating with one another. We find that one person shows high consistency with LLM labels, with an overlap of 0.76. Their majority-voted label across annotators shows an overlap with LLM labels of 0.70. These results show the reliability of LLMs in annotating text data.

C Results using other LLMs are consistent

Across the paper, we intentionally adopted models from the same company (OpenAI) because we wanted the entire process to be shaped by a consistent set of cultural biases, making it more straightforward to expose and analyze these biases.

To increase the generalizability of our work, we explored other popular open-sourced or closed models, including Deepseek-chat-v3-0324 [5], Mistral 24b instruct [6], and Gemini 2.5 Flash-Lite [7], to replicate the full set of experiments. Overall, we find our conclusions to be robust. Researchers have observed increasing convergence among LLMs, largely because major model families are trained on similar mixtures of large-scale web corpora, code repositories, and curated texts (Wu et al., 2025; Wolfram and Schein, 2025). If several LLMs from different companies independently articulate the same belief-change dynamics, this convergence itself may be informative for the biases learned from in-the-wild, human-generated corpora. Specifically:

- Alignment between priors and human comments, but not posteriors: All three additional models show a relatively consistent set of priors as OpenAI models used in the main paper, and align with human comments about the definition of "good science", compared to the low alignment for posteriors VS human comments. Human reviewers tend to explicitly reward aspects that moderately align with LLMs' normative priors: OpenAI correlation = 0.49; Deepseek 0.28; Gemini 0.65; Mistral 0.45. By contrast, human reviewers generally avoid articulating posteriors in their review comments (OpenAI correlation = - 0.14; Deepseek = –0.16; Gemini = -0.20; Mistral = -0.22), but all models do articulate these posteriors.

- Shift from prior to posterior: During the transition from prior to posterior, we still observe a systematic shift from rigor-related aspects toward storytelling perspectives. All models demonstrate surprisingly high consistency in using certain heuristics such as "whether the paper tells a good interdisciplinary story", and across models, they show on average 89% agreement with Open-AI's judgment. This hypothesis also predictably gained the highest prior-to-posterior attention gain: interdisciplinary integration (OpenAI top-1; DeepSeek top-1; Gemini top-2; Mistral top-5) and contextualization (DeepSeek top-2; OpenAI top-2; Gemini top-1; Mistral top-1) are both systematically overestimated in the posterior belief shift. By contrast, rigor is consistently underestimated (OpenAI lowest top-1; DeepSeek lowest top-1; Gemini lowest top-2; Mistral lowest top-2).

D Out-of-sample judgments

For robustness, we experimented with another set of 5,000 pairs. We find the generated posteriors in the main paper can still explain a substantial proportion (>95%) of pairs, even though the models had not seen them at all. Moreover, these posterior distributions are highly correlated (0.80) with those on the 5000 pairs in the original paper. We also experimented with using LLMs to generate hypotheses from the new dataset. We found that the generated hypotheses are qualitatively consistent, with only minor variations in order. We briefly noted this in

[5] https://api-docs.deepseek.com/news/news250325

[6] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

[7] https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite. The result of Gemini 2.5 Flash is very similar to that of its lite version.

the original paper (see Section 4 iterative search). We conclude that scientific judgments can often be explained by a compact set of generalizable hypotheses—spanning both easy heuristics and harder-to-discern, quality-driven priors—reflecting recurring evaluative patterns across diverse cases.

E Fidelity experiments

Summarization using methods from (Zhang et al., 2025) demonstrates high fidelity compared to the raw text. In Figure 7, panel (a), we use the embedding model from (Singh et al., 2023), specifically pretrained on scientific papers' citation relations using contrastive learning. We embed both the summarizations and their corresponding raw texts in each section. To visualize and compare the distribution of these embedding areas, we reduce their dimensionality to two. We then estimate the boundary of each region using a 5% Gaussian kernel density estimate, drawing contours that enclose the densest 95% of data points in each embedding set. If the summarizations contain significant hallucinations, we would expect their embeddings to cover meaningfully different regions than those of the raw texts. However, in panel (a), we find that the distributions largely overlap, suggesting strong semantic alignment. In panel (b), we compute the cosine similarity between the two (raw, summarized) embeddings and find that most values are above 0.95.

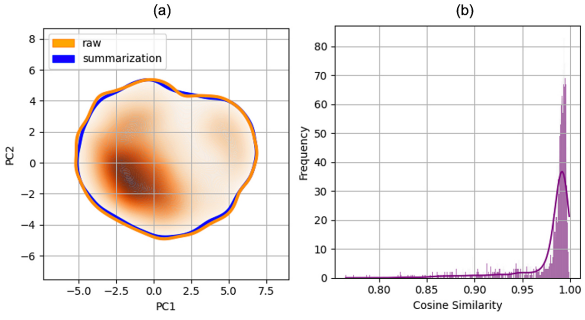


Figure 7: The summarization exhibits minimal hallucination (panel a - almost overlapping embeddings) and a high degree of fidelity (panel b).

We conducted a complementary robustness check to assess whether LLMs make consistent judgments when directly reading full texts, compared to the extended summarization used in the paper for producing generalizable hypotheses under the limited context window. If using full texts from parsed PDFs, the results are essentially identical to those obtained with extended summaries (given their high fidelity), yielding a consistency score of 0.89 (i.e., the same judgment) across 1,000 judgments.

Algorithm 1: Hypothesis Search

Initialize an empty hypothesis set H and unexplained set W (initially all pairs);
for *each time step* t **do**
  Generate 5 hypotheses on top of the existing set H, from 50 random pairs of papers within the unexplained set W;
  Update hypothesis set H;
  Apply 5 new hypotheses to all pairs in W by 3-fold confidence-weighted voting;
  Apply 5 new hypotheses to explained pairs by 3-fold confidence-weighted voting as well;
  Update the unexplained sample set W;
  if $W < 5\%$ *of all samples* **then**
    Stop searching;
  end
end
Compare searched hypotheses with LLM priors extracted by self-talk;

F Overall algorithmic framework

The overall algorithmic framework is shown in Algorithm 1.

G Related work

Hypothesis Generation: Recent studies have explored the capabilities of LLMs in generating novel, testable, and interpretable hypotheses across a range of domains. Zhou, Liu et al. (Liu et al., 2024a, 2025; Zhou et al., 2024) proposed a framework that prompts LLMs to generate hypotheses from labeled datasets, beginning with a few examples. These initial hypotheses are iteratively refined using a reward mechanism inspired by the upper confidence bound algorithm from multi-armed bandits, balancing exploration and exploitation. Other works have integrated external knowledge — such as scientific literature, knowledge graphs, and raw web corpus —

into the hypothesis generation process (Xiong et al., 2024; Kulkarni et al., 2025; Wang et al., 2023, 2024; Yang et al., 2024), improving the accuracy and novelty of LLM-generated hypotheses while reducing hallucinations. This line of work demonstrates the potential of LLMs in advancing scientific discovery. Beyond computer science, other disciplines are also beginning to explore these possibilities. Ludwig and Mullainathan (Ludwig and Mullainathan, 2024) highlighted the promise of LLMs as tools for hypothesis generation in economics, arguing that machine learning models can uncover patterns and relationships that traditional econometric methods may miss. Manning, Zhu, and Horton (Manning et al., 2024) presented a simulation-based approach for automatically generating and testing social science hypotheses in silico. In each simulated agent-interaction case, causal relationships are both proposed and tested by the LLM system, with results aligning well with those derived from economic theory, particularly in predicting the signs of estimated effects. In medicine, LLMs have been used to hypothesize synergistic drug combinations for breast cancer treatment, with several predictions later validated in laboratory experiments (Bazgir et al., 2025), showcasing the potential of LLMs to surface effective, previously overlooked therapeutic strategies.

Implicit Bias in LLMs: Concerns about bias in AI systems have long focused on the ways these technologies can absorb and amplify harmful content from large-scale human-generated datasets – particularly prejudices related to race, gender, and other social dimensions (Ntoutsi et al., 2020). Large language models (LLMs), trained on vast corpora of human text, are especially prone to perpetuating and even intensifying these biases (Weidinger et al., 2021; Gallegos et al., 2024). Such biases are not computational artifacts but reflect the deeper cultural and social frameworks that shape human communication (Babaeianjelodar et al., 2020; Lee et al., 2024a; Guilbeault et al., 2024). The recognition that LLMs encapsulate and propagate societal biases and heuristics is now well-established in literature, framing these models as inherently cultural and social technologies (Blodgett et al., 2020; Farrell et al., 2025; Weidinger et al., 2021; Naous et al., 2023; Li et al., 2024a). Such representations have also been used to study culture itself (Caliskan et al., 2017; Garg et al., 2018; Kozlowski et al., 2019), and subtler biases can be probed indirectly (Chen et al., 2025; Fang et al., 2024); we extend this diagnostic stance to prior-to-posterior shifts. Much of the existing research on LLM bias has concentrated on explicit, easily identifiable forms of bias – such as derogatory language, disparate system performance, erasure, exclusionary norms, misrepresentation, stereotyping, and toxicity (Gallegos et al., 2024; Czarnowska et al., 2021). Meanwhile, efforts in AI alignment, including outer alignment, inner alignment, and interpretability, have sought technological strategies for migrating such risks (Shen et al., 2023; Meade et al., 2021; Kumar et al., 2022). Despite progress in addressing explicit bias, unwritten codes in LLMs' "cognitive" process remain understudied. Like human cognition, LLMs may exhibit subtle, unconscious forms of bias even when explicit prejudices are absent (Shen et al., 2023; Lin and Li, 2025). These implicit biases are often harder to detect yet can have deeper and more pervasive effects on model outputs and downstream decision-making. Thus, systematically identifying these "unwritten codes" in LLMs is crucial to ensuring fairness, accuracy, and trustworthiness (Lin and Li, 2025). One method worth mentioning in this topic is (Bai et al., 2025). The authors draw on experimental paradigms from the Implicit Association Test (IAT) in psychology to design prompts that reveal latent biases in LLMs. Rather than directly asking models about their views on particular social groups – which, like humans, they often avoid answering – they instead present structured tasks that require associations or comparative decisions. For instance, a model might be asked to assign positive or negative words to names with racial or gender connotations, or to choose which of two candidates is more suitable for a specific role. This prompt-based approach bypasses explicit self-reports and mirrors the logic of involuntary responses in psychological testing, allowing underlying implicit stereotypes to surface indirectly through model behavior.

Science, Innovation, and LLM-as-a-Judge: Prior work has increasingly applied LLMs to assist researchers by generating feedback, reviews, or reflections throughout the research process (Yuksekgonul et al., 2024; Janssen et al., 2023). Most recent studies have focused on addressing challenges posed by the rapid growth of academic publishing, such as

rising review workloads and declining review quality (Azad and Banu, 2024; Rogers et al., 2023). Liang et al. (Liang et al., 2024) evaluate GPT-4's capacity to generate scientific feedback and find substantial overlap with human reviews, especially in early manuscript stages. Thakkar et al. (Thakkar et al., 2025) developed a "Review Feedback Agent" combining multiple LLMs; in a randomized trial, 27% of reviewers revised their feedback after receiving LLM suggestions, which were more detailed and informative. Tyser et al. (Tyser et al., 2024) propose a broader framework, training LLMs to align with human preferences through pairwise comparisons and enhancing review robustness with techniques like adaptive prompting and role-playing. While prior studies on LLMs in peer review have primarily focused on augmenting or replicating human judgment to address scalability and quality concerns, our argument departs from this substitutional framing. Instead of merely simulating reviewers to reduce workload or enhance coverage of aspects, we leverage LLMs' generative capacity to surface the implicit heuristics and tacit values that often shape peer review decisions but remain unspoken. By prompting LLMs to hypothesize why one paper is rated more highly than another, we reveal not only individual reviewer preferences but also broader cultural norms and evaluative codes that govern scientific judgment. In doing so, we treat AI not as a replacement for human thinking but as a diagnostic tool – one that can reflect, exaggerate, and ultimately make visible the "unwritten code" underpinning peer review and, more broadly, other sociotechnical systems.

H Details of the dataset of conference submissions

This dataset includes 26,731 submissions to 46 conferences: 1st ContinualAI Unconference, AAAI Conference on Artificial Intelligence 2024 and 2025, Symposium on Advances in Approximate Bayesian Inference 2024, Conference on Language Modeling 2024, Cooking Robotics Workshop 2024, Conference on Robot Learning 2023 and 2024, Conference on Parsimony and Learning 2024, IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 and 2024, Workshop on Distributed Infrastructure for Common Good 2023, Workshop on Embodiment-Aware Robot Learning 2024, European Conference on Computer Vision 2024, Conference on Empirical Methods in Natural Language Processing 2023, European Space Power Conference 2023, Fast, Low-resource, and Accurate Organ and Pan-cancer Segmentation in Abdomen CT 2023, SIGIR Workshop on Generative Information Retrieval 2024, ACM/IEEE International Conference on Human-Robot Interaction 2023 and 2024, International Conference on Integration of Science and Technology for Sustainable Development 2024, International Conference on Learning Representations 2017 to 2025, International Conference on Machine Learning 2023 and 2024, ACM International Conference on the Theory of Information Retrieval 2024, International Joint Conference on Artificial Intelligence 2024, International Semantic Web Conference 2024, ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2023 and 2024, ACM International Conference on Multimedia 2024, Conference on Neural Information Processing Systems 2021-2024, Neuro-Symbolic Learning and Reasoning in the era of Large Language Models 2024, Next-generation Data Governance Workshop 2024, Tsinghua University Advanced Machine Learning 2024.

Each paper in the dataset is represented with a structured format while retaining details. Each part in the representation is generated by feeding the corresponding sections into the LLM.

Title:
Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness

Raw abstract:
Among a wide range of success of deep learning, convolutional neural networks have been extensively utilized in several tasks such as speech recognition, image processing, and natural language processing, which require inputs with large dimensions. Several studies have investigated function estimation capability of deep learning, but most of them

have assumed that the dimensionality of the input is much smaller than the sample size. However, for typical data in applications such as those handled by the convolutional neural networks described above, the dimensionality of inputs is relatively high or even infinite. In this paper, we investigate the approximation and estimation errors of the (dilated) convolutional neural networks when the input is infinite dimensional. Although the approximation and estimation errors of neural networks are affected by the curse of dimensionality in the existing analyses for typical function spaces such as the Holder and Besov spaces, we show that, by considering anisotropic smoothness, they can alleviate exponential dependency on the dimensionality but they only depend on the smoothness of the target functions. Our theoretical analysis supports the great practical success of convolutional networks. Furthermore, we show that the dilated convolution is advantageous when the smoothness of the target function has a sparse structure.

Contextual background:
Convolutional Neural Networks (CNNs) are widely used in applications like image recognition, speech processing, and natural language understanding. These applications often involve inputs that are extremely high-dimensional—so high, in fact, that they can be treated as infinite-dimensional in theory. However, most theoretical research on neural networks has focused on low- or fixed-dimensional settings, assuming the input has fewer dimensions than the number of samples available for training. Existing theoretical results on neural networks, such as those studying their function approximation capabilities, break down or become less meaningful in infinite-dimensional cases. Some prior studies attempted to address this by assuming inputs lie on a low-dimensional manifold or that the target function has certain regularity properties. Still, none fully addressed the behavior of CNNs when inputs are truly infinite-dimensional. This paper fills that gap. It explores whether CNNs can still learn effectively in such extreme settings by assuming the target function behaves differently across input dimensions, a property called anisotropic or mixed smoothness.

Key idea:
The main idea of the paper is that CNNs, especially those using dilated convolutions, can effectively learn even when the input is infinite-dimensional, as long as the target function has a particular kind of smoothness across dimensions. Rather than relying on the number of input dimensions, the authors argue that how the target function changes across different input coordinates is what truly determines learnability. This is formalized using two smoothness conditions: Mixed smoothness, where the function may vary differently across each input direction but still follows a structured form. Anisotropic smoothness, where the function is especially smooth (or insensitive) in certain input directions and more complex in others. The paper introduces a formal function space to describe such functions and then proves that CNNs can achieve good learning performance in these spaces, without being affected by the sheer dimensionality of the input. The key is that the CNNs, particularly with dilated convolutions, are capable of automatically detecting and focusing on the most relevant input components.

Theoretical and methodological details:
This section contains the core technical contributions of the paper. A. Learning Problem The paper studies a nonparametric regression problem, where the goal is to learn a function that maps an infinite-dimensional input

to a real-valued output. The inputs can be thought of as sequences or functions, such as raw waveforms, high-resolution images, or long text embeddings. The challenge is to characterize how well a neural network can approximate this unknown function and how much error we can expect when training from a finite number of samples. B. Smoothness-Based Function Space To handle infinite-dimensional inputs, the authors define a new function space where functions are categorized based on their smoothness. Crucially, this smoothness is allowed to vary across input coordinates: Some directions may be very "rough" (i.e., important to capture precisely), while others can be ignored or treated coarsely. This idea mirrors how images contain both low-frequency (broad structure) and high-frequency (detail) components, and not all are equally important. By introducing a flexible way to quantify how much a function varies in each input direction, the authors build a mathematical foundation for analyzing the performance of neural networks in these spaces. C. Approximation Capabilities The paper first analyzes fully connected neural networks and shows that they can approximate functions in the proposed smoothness space well—but only if the network is explicitly given the most important input dimensions. In other words, these networks are theoretically powerful but need help in selecting relevant inputs. In contrast, CNNs can discover the important directions automatically thanks to their structured and hierarchical nature. The authors show that: Ordinary CNNs perform well when the smoothness of the target function increases gradually across input directions (e.g., the first few coordinates are rough, the rest smooth). Dilated CNNs are even more powerful when the target function has a sparse structure—that is, only a few input directions matter significantly, and these may be far apart. The authors prove that in both settings, CNNs can achieve fast convergence rates—meaning the prediction error decreases quickly as more data is observed—even if the input is infinite-dimensional. These rates depend only on the smoothness of the function, not on the number of input dimensions. D. Estimation from Data The paper also analyzes how well neural networks trained on data (using empirical risk minimization) can learn the true function. They show that: CNNs achieve dimension-independent learning rates under the right smoothness assumptions. For functions with gradually increasing smoothness, shallow CNNs suffice. For sparsely smooth functions, deeper or dilated CNNs are needed to identify and capture distant important features. This theoretical result aligns well with practical intuitions about CNNs being good at capturing both local and global structures.

Experiments and results:
The paper does not include any empirical experiments—no training on real datasets, no benchmarks, no runtime results. Instead, it provides a theoretically rigorous analytical framework to evaluate model performance, focusing on: Approximation error: How well the neural network class can represent the true function. Estimation error: How close the trained model gets to the true function when trained on finite data. The authors derive explicit mathematical expressions for the convergence rates of neural networks. These rates are shown to be: Polynomial in the number of samples, not exponential in the input dimension. Optimal or near-optimal, matching or extending existing bounds in lower-dimensional settings. They also compare their rates with those from other methods like kernel regression or traditional non-deep learning techniques, demonstrating that CNNs offer better or comparable rates in infinite-dimensional settings—especially when function smoothness is sparse or non-uniform. This analytical evaluation shows

that CNNs, especially those using dilated convolutions, are theoretically justified for high- and infinite-dimensional data, even in the absence of experimental validation.

Mentioned impact:
This paper makes a theoretical contribution by offering comprehensive analyses of deep learning's learnability in infinite-dimensional spaces. Its results: Support the practical success of CNNs in real-world tasks involving high-dimensional data. Justify the use of dilated convolutions as an effective tool for handling sparse or long-range dependencies. Advance our understanding of when and why deep learning avoids the curse of dimensionality, shifting the focus from input size to function structure.

I Prompts used in the paper

#Hypothesis evaluation
You are a computer scientist.
Task: Evaluate the following hypothesis based on the content of two papers.
Hypothesis: "{hypothesis}"
Paper 1 Content: {row['comprehensive content 1']}
Paper 2 Content: {row['comprehensive content 2']}
Instructions:
1. Analyze the content from both papers carefully to determine whether Paper 2 is superior in the aspects identified in the hypothesis.
2. Provide your response in the exact JSON format (without additional commentary):

```
<label>X</label>


<confidence>Y</confidence>
```

- X: "1" if Paper 2 is superior in the identified aspects (i.e., the comparison of contents supports the hypothesis), or "0" if not.
- Y: An integer between 0 and 10 indicating your confidence level in this judgment (with 10 meaning highest certainty).

#Hypothesis iterative search
You are a computer scientist.
You are analyzing two papers submitted to the same venue.
Paper 1 (comprehensive_content_1) received a lower peer review score (rating_1) than Paper 2 (comprehensive_content_2 and rating_2).
Your task is to generate 5 specific, creative, and critical hypotheses explaining why Paper 1 may have received a lower score.

- Base your hypotheses on a rigorous content-driven comparison of the provided dataset df, identifying concrete differences that might have influenced peer review scores.
- Avoid vague reasoning and ensure each hypothesis is distinct, well-reasoned, and actionable.
- Ensure that the generated hypotheses are distinct from those in df_hypothesis.

Provide your response in the exact JSON format (without additional commentary):
"Compared to another paper, one paper [your hypothesis]."

#Prior hypothesis search
You are a computer scientist.
Task: Generate specific, creative, and critical hypotheses about concrete differences between two papers that might have influenced their peer review scores. You will perform 4 rounds. In each round, generate 5 hypotheses.

Each hypothesis must:
- Focus on a distinct, well-defined, and actionable factor.
- Avoid vague reasoning.

Provide your response in the exact JSON format (without additional commentary):
"Compared to another paper, one paper [your hypothesis]."

#Annotation of human review comments
#attach examples provided to LLMs for three conditions across 20 hypotheses
You are a computer scientist.
Please annotate the human feedback below by mapping it to the aspects in the 20 hypotheses.
Return your answer in exactly the following JSON format (no extra text or commentary):

```
{"scores": [X1, X2, X3, ..., X20]}
```

Each Xi must be one of: `-1`, `0`, or `1`
- Use `1` if the feedback praises the standard
- Use `-1` if the feedback criticizes the standard
- Use `0` if the standard is not mentioned

Hypotheses:
`{hypothesis_list}`

Feedback:
`{feedback_text}`

J The list of generated hypotheses

The list of generated hypotheses is shown in Table 1.

K Cross-validation dictionary

The dictionary method used in the paper to complement LLM labeling/matching is shown in Table 2. Note that this heuristic dictionary-like approach, while widely used in social science (Guo et al., 2016; Frankel et al., 2022), ultimately serves as a substitute rather than the major methods of labeling human comments of "mention" or matching priors and posteriors, presented in the paper.

L Humans explicitly reward the internal qualities of science, but a significant hidden driver remains.

We use GPT-4o with few-shot to annotate human review comments across the aspects of 20 hypotheses. Each aspect is labeled as praise (1), not mentioned (0), or criticism (-1). We then run an OLS regression (using `scipy.stats` package) to examine how reviewers' attitudes toward these aspects relate to the scores they assign within the conference. The results are shown in Table 3.

We have two points to discuss:

- Only five hypotheses are statistically significant meaningful predictors of human review scores, and all have positive coefficients—indicating that when these aspects mentioned in the hypothesis are positive, the associated paper tends to receive a higher score. These hypotheses (highlighted in red) clearly reflect the internal qualities of scientific work. This suggests that reviewers do use their scores to explicitly reward normative aspects of science. Notably, two of these five hypotheses—theoretical rigor and the balance between theoretical analysis and practical impact—also appear among the top five hypotheses favored by priors but not favored by posteriors.

- However, the regression model's explanatory power is very low ($R^2 = 0.06$). This suggests that much of what drives reviewers' scoring decisions is not captured by the explicit reward of these normative qualities. There are implicit factors strongly influencing the review process that are not reflected directly in the scores.

M Discussions of position bias in LLM-as-a-judge

To address potential position bias, we adopt two mitigation strategies in the original paper. First, we construct paper pairs with substantial quality differences, as prior research has shown that position bias is most pronounced when the quality gap is small (Li et al., 2024b; Shi et al., 2024). Second, we randomize the positions of Paper 1 and Paper 2 across voting, while ensuring logically consistent prompts and vote aggregation. We then assess whether position bias is indeed stronger when the quality signal gap is small. Specifically, we regress the consistency rate — defined as the proportion of consistent judgments across 20 hypotheses after swapping paper positions — on the review score gap between the two papers. Our analysis (Table 4) shows that a larger review score gap is indeed associated with greater consistency, and the effect is statistically significant, though weak in magnitude.

Hypothesis	Round	Prior	Posterior	Rank in shifts	Mention
lack justification regarding its novelty	1	0.36	0.20	13	0.67
not theoretically rigorous	1	0.92	0.18	20	0.57
the design is over-engineered, not elegant, and unnecessarily complicated	1	0.50	0.02	18	0.74
not use a fair benchmark for evaluation	1	0.32	0.08	15	0.16
lack implementation and reproducibility details	1	0.44	0.18	16	0.05
lack comprehensive ablation and hyperparameter analysis	2	0.58	0.35	14	0.26
the contribution is incremental	2	0.06	0.24	9	0.12
not balance theoretical justification and practical impact	2	0.68	0.20	17	0.60
the presentation is too complicated	2	0.14	0.50	4	0.45
unclear about its contextualization within the related literature	2	0.30	0.65	5	0.50
misaligned with the emerging scope	3	0.14	0.10	11	0.11
lack empirical evidence in terms of model and architecture design choice	3	0.16	0.51	6	0.18
limited in scalability to real-world scenarios	3	0.44	0.35	12	0.20
insufficient in testing robustness such as uncertainty and bias in evaluation experiments	3	0.84	0.28	19	0.53
insufficiently discuss social and ethical aspects of its algorithms	3	0.18	0.55	3	0.02
weakly motivated in its training and optimization techniques	4	0.30	0.62	7	0.82
the theoretical assumptions are unrealistic	4	0.40	0.58	8	0.61
poorly explore edge cases and failure modes	4	0.58	0.60	10	0.24
require dense prior background knowledge	4	0.02	0.59	2	0.07
not interact well with adjacent domains, such as statistics and cognitive science	4	0.06	0.77	1	0.01

Table 1: The generated hypotheses. A shift in rank indicates the change in "importance", measured as posterior minus prior (i.e., a high shift means the hypotheses that were not highly weighted in the prior but were used much more frequently in actual judgments). Orange highlights the top 5 hypotheses with the largest gains (e.g., contextual and narrative perspectives), while blue highlights the top 5 with the largest losses (e.g., theoretical rigor and normative perspectives). "Mention" indicates the proportion of human comments that mentioned the hypothesis, as identified by GPT-4o. All hypotheses start with "Compared to the other, one paper..."

Hypothesis Index	Dictionary
1	["novelty", "novel"]
2	["rigor", "rigorous", "theoretically rigorous", "theoretical rigor", "proof"]
3	["over-engineering", "over-engineered", "elegance", "elegant", "complicated model", "complicated architecture", "clarity"]
4	["fair benchmark", "benchmarking", "evaluation"]
5	["reproducibility", "implementation detail"]
6	["ablation", "hyperparameter"]
7	["incremental", "A+B"]
8	['theoretical justification', "practical impact"]
9	["presentation", "readability", "writing"]
10	["related work", "contextualization", "literature"]
11	["emerging", "out of date", "scope", "trend"]
12	["design choice", "empirical evidence", "architecture design", "model design"]
13	["scalability", "real-world", "applicability"]
14	["uncertainty", "biased evaluation", "evaluation bias", "evaluation", "robustness"]
15	["societal", "ethical", "society", "ethic", "social"]
16	["optimization", "training", "train", "loss function"]
17	["theoretical assumption", "practical", "practically", "in practice", "unrealistic assumption"]
18	["failure mode", "edge case"]
19	["implicit knowledge", "prior knowledge", "implicit background", "prior background", "background knowledge"]
20	["interdisciplinary", "multidisciplinary", "statistics", "cognitive science", "neural science", "psychology", "cognition"]

Table 2: The dictionary method used in the paper to complement LLM labeling/matching.

Table 3: Regression results on the mentioned perspectives and human scoring.

Hypothesis	Coef.	Std. Err.	t	P>|t|	[0.025, 0.975]
const	5.966	0.100	59.508	0.000	[5.769, 6.163]
lack justification regarding its novelty	0.231	0.096	2.410	0.016	[0.043, 0.420]
not theoretically rigorous	0.209	0.095	2.194	0.028	[0.022, 0.396]
the design is over-engineered, not elegant, and unnecessarily complicated	0.046	0.100	0.463	0.643	[-0.150, 0.242]
lack comprehensive ablation and hyperparameter analysis	0.152	0.087	1.754	0.080	[-0.018, 0.322]
unclear about its contextualization within the related literature	0.069	0.082	0.836	0.403	[-0.093, 0.230]
the contribution is incremental	0.205	0.089	2.310	0.021	[0.031, 0.379]
not use a fair benchmark for evaluation	0.169	0.085	1.998	0.046	[0.003, 0.336]
the presentation is too complicated	-0.006	0.081	-0.078	0.938	[-0.166, 0.153]
lack implementation and reproducibility details	0.019	0.098	0.198	0.843	[-0.172, 0.211]
not balance theoretical justification and practical impact	0.356	0.136	2.619	0.009	[0.089, 0.622]
misaligned with the emerging scope	0.193	0.223	0.865	0.388	[-0.245, 0.630]
insufficiently discuss social and ethical aspects of its algorithms	-0.051	0.262	-0.196	0.844	[-0.565, 0.462]
not interact well with adjacent domains, such as statistics and cognitive science	-0.081	0.249	-0.325	0.745	[-0.569, 0.408]
lack empirical evidence in terms of model and architecture design choice	0.015	0.102	0.149	0.882	[-0.185, 0.215]
limited in scalability to real-world scenarios	0.034	0.110	0.308	0.758	[-0.183, 0.251]
insufficient in testing robustness such as uncertainty and bias in evaluation	0.026	0.204	0.127	0.899	[-0.374, 0.426]
weakly motivated in its training and optimization techniques	0.073	0.169	0.430	0.668	[-0.259, 0.404]
the theoretical assumptions are unrealistic	0.041	0.175	0.232	0.817	[-0.302, 0.383]
poorly explore edge cases and failure modes	-0.030	0.080	-0.374	0.708	[-0.186, 0.127]
require dense prior background knowledge	0.006	0.099	0.056	0.955	[-0.189, 0.200]

Table 4: OLS regression results on the position bias. Standard errors assume that the covariance matrix of the errors is correctly specified.

Dep. Variable:	consistency	**R-squared:**	0.015
Model:	OLS	**Adj. R-squared:**	0.015
Method:	Least Squares	**F-statistic:**	77.26
Prob (F-statistic):	2.03e-18	**Log-Likelihood:**	5245.2
No. Observations:	5000	**AIC:**	-1.049e+04
Df Residuals:	4998	**BIC:**	-1.047e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.74	0.00	237.00	0.00	0.73	0.75
gap	0.01	0.00	8.79	0.00	0.01	0.01

Omnibus:	18.990	**Durbin-Watson:**	1.985
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16.480
Skew:	-0.085	**Prob(JB):**	0.000264
Kurtosis:	2.777	**Cond. No.:**	9.08