# DOGE: Reforming AI Conferences and Towards a Future Civilization of Fairness and Justice

Zeyuan Allen-Zhu
zeyuan2025@allen-zhu.com
Allen-Zhu Research

Xiaoli Xu
xiaoli2025@allen-zhu.com
Allen-Zhu Research

February 2, 2025

(full version 1)

**Pumping Elephant**[1]

AI conferences are expanding rapidly, yet the quality of peer reviews are declining. Authors risk damage to their reputation for mistakes, while anonymous reviewers can escape accountability if they make irresponsible judgments. In this paper, we propose DOGE 1.0, a system that uses *AI-models to arbitrate* reviewer-author disputes. We explain why AI arbitration could mitigate issues including human bias, emotional or malicious behavior, and present criteria for assessing whether current AIs are *sufficiently intelligent* for this role.

By classifying "intelligence" into four levels (L1–L4), we argue that *arbitrating* disputes (L1) requires far less intelligence than *authoring* a paper (L4), *reviewing* it (L3), or even *auditing* a review (L2). For AI conferences, our evidence shows that many state-of-the-art AIs **already excel** at L1 tasks, though **only a few** approach L2-level. We therefore propose employing AIs with reliable L1 (ideally L2) intelligence as neutral arbitrators. For theoretical computer science experts, this may be unsurprising: a polynomial-time verifier in $\mathsf{P}$ can validate languages in $\mathsf{NP}$, or even $\mathsf{PSPACE}$ via interactive proofs, demonstrating that even a relatively weak arbitrator can effectively help assess L4-level work.

We also hint at DOGE 2.0, which could introduce robust, persistent incentives for reviewers — potentially including crypto-based rewards or penalties — to foster a fairer and more accountable peer-review system. Moreover, the theory behind the L1–L4 intelligence hierarchy provides criteria for determining when AI models have attained the *necessary* intelligence to serve as neutral arbitrators in any field, potentially promoting fairness and justice in our civilization — or at the very least, adding a dash of fun back to AI conferences.

## 1 Introduction

> "The arc of the moral universe is long, but it bends toward justice." — Martin Luther King Jr.

ICML, ICLR, and NeurIPS are arguably the three most prestigious AI conferences. In recent years, these conferences have seen exponential growth in submissions. However, the quality of reviews has noticeably declined. Although some organizers have introduced measures — such as the

[1] "Pumping Elephant" was an accidental mistranslation of the Chinese word " 抽象 (abstract)" in a real published paper under a "compromised peer-review process" (their original phrase) at Springer [27].

Project page (with missing materials): http://doge.allen-zhu.com.

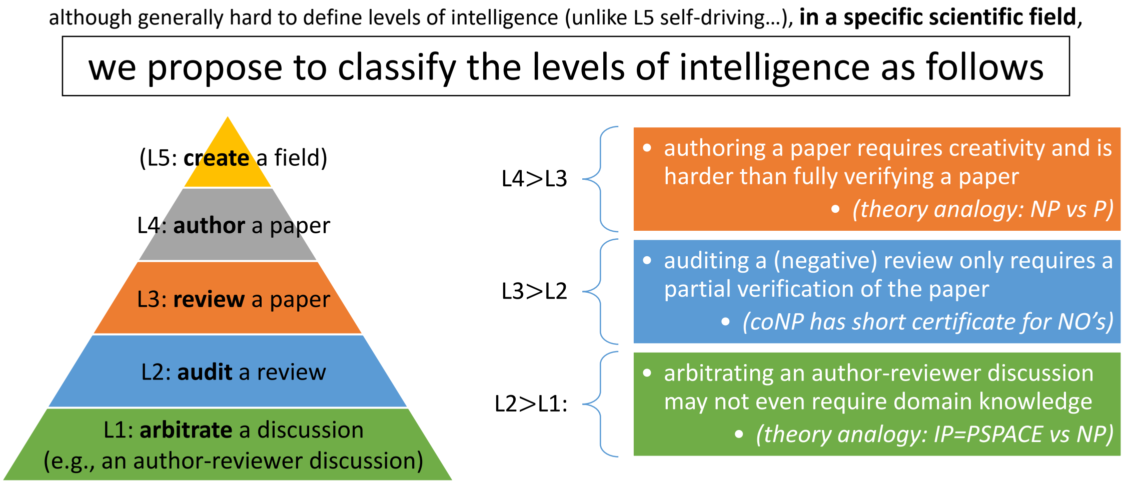



Figure 1: Our proposal on a categorization of "intelligence" in a given scientific field.

ICLR 2025 policy requiring authors to review other submissions — it remains unclear whether these steps alone can sufficiently address the challenges brought on by the rapidly expanding community.

While much of the discussion centers on whether reviewers use AI, we focus on the issues *caused by human reviewers* and their harmful impacts. Although many reviewers act in a relatively responsible manner, inherent vulnerabilities in the process allow issues to surface. Our investigation aims to design a new, more efficient reviewing system that mitigates these human-driven issues and, in doing so, restores AI conferences to their former glory.

Besides the standard concerns regarding reviewers' quality decline (I0), in Section 2 we detail six additional issues plaguing the current review system: (I1) anonymity permits irresponsible critiques, (I2) human stubbornness leads to unrevised errors, (I3) emotional influence may bias evaluations, (I4) malicious behavior can distort assessments, (I5) the lack of real incentives discourages quality review, and (I6) senior chairs often hesitate to override flawed assessments. These factors can lead to random or unfair decisions, stress authors, and ultimately reduce the credibility of AI conferences.

> If this trend continues, peer-reviewed AI conferences may become *increasingly awkward*. With social media platforms like Twitter, YouTube, and GitHub emerging as unstoppable forces for promoting work, the traditional role of conferences in establishing credibility is under immense pressure.

Our proposed solution relies on a categorization of "intelligence" (see also Figure 1):

- **L4:** the intelligence needed to *author* scientific papers,
- **L3:** the intelligence to *review* scientific papers,
- **L2:** the intelligence to *audit* the reviews of scientific papers,
- **L1:** the intelligence to *arbitrate* a reviewer-author discussion about a scientific paper.

In Section 3, we provide theoretical evidence suggesting that, in any scientific field, a hierarchy emerges as $L4 > L3 > L2 > L1$. For AI conferences, we present a real-life example — our own ICLR 2025 submission, which was rejected by a meta-reviewer (a.k.a. area chair, AC) based on three false claims — to demonstrate that some currently available AI chatbots appear to have achieved at least L2-level intelligence in this context. Specifically, models such as Gemini 2.0 Flash Thinking, OpenAI o1, and DeepSeek R1 were able to effectively identify those three mistakes without any author input, whereas other models only reached L1-level intelligence, meaning an author has to guide them to where the errors were.

Based on these observations, in Section 4 we propose the "DOGE 1.0 protocol," an *AI-based arbitrator system* tailored for the peer-review process at AI conferences. In the traditional system, a reviewer raises concerns and the author responds to convince each other — often over multiple rounds. In contrast, the AI-based arbitrator system employs a chatbot (with at least L1-level, ideally L2-level, intelligence) to arbitrate the discussion. In short:

> The author and reviewer now *seek to convince the arbitrator* via multi-round conversation of whether the paper should be accepted or rejected.

The core idea is that the arbitrator does not need the sophistication of L3 or L4 intelligence; it does not even need L2 intelligence — it only needs to follow basic logic and fact checking to evaluate the claims made by both sides.

Section 4 further explains the advantages of the DOGE 1.0 protocol and shows how it can fully or partially mitigate the seven issues (I0–I6) noted earlier. In brief, DOGE 1.0 discounts low-quality reviews (addressing I0) and leverages AI's logical reasoning to replace human reviews that are stubborn, emotional, or even malicious (addressing I2–I4). It further enables models to speak up without fear of reprisal (addressing I6), employs public and upgradeable AI arbitrators to overcome the anonymity issue (addressing I1), and uses tireless AI arbitrators that do not require incentives to ensure consistent review quality (addressing I5).[2]

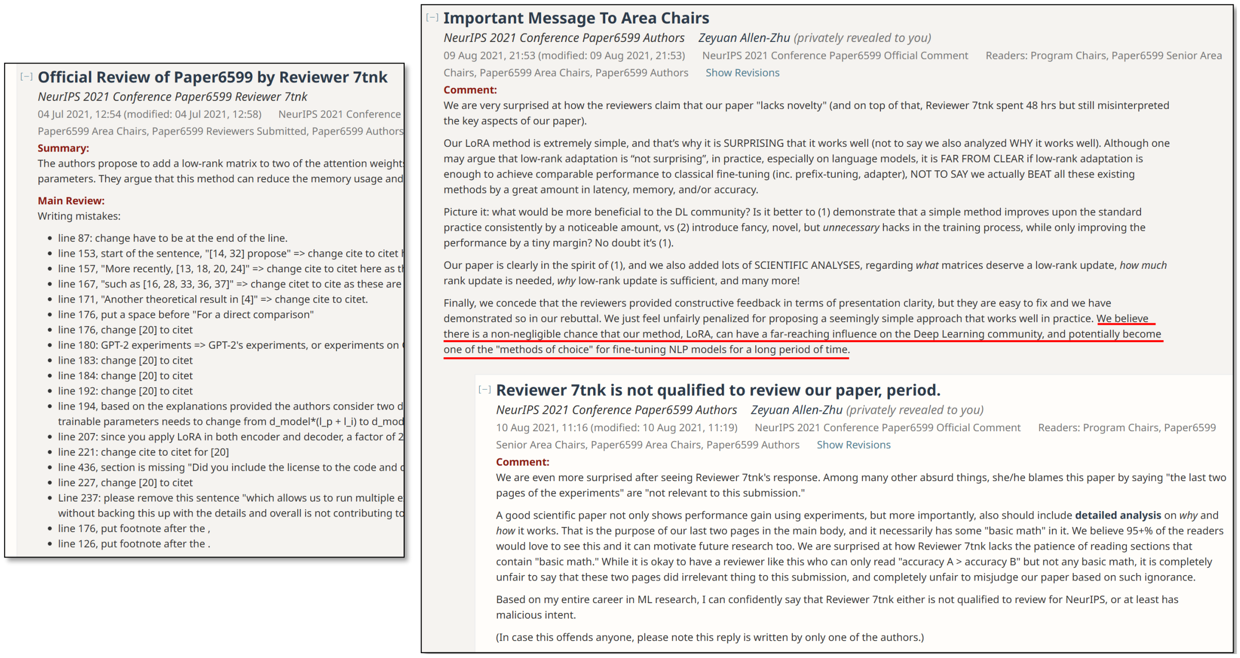
In Section 5, we explore the potential evolution to a "DOGE 2.0" system. This enhanced system would incorporate additional incentives for human reviewers to submit higher-quality reviews — and would, crucially, impose "punishments" when they fail to do so. Today, authors bear real reputational risks when mistakes surface in their papers, whereas anonymous reviewers face no comparable consequences. We discuss how an AI arbitrator could estimate reviewer quality and publicly record global scores for reviewers, potentially integrating privacy-preserving crypto coins specific to AI conferences to reward or penalize them accordingly.

Finally, in the Conclusion we reflect on how far we remain from reaching L3-level intelligence and propose new benchmarks for its evaluation. We also highlight the crucial role of senior human experts in the field in engaging with and better educating the AI-based arbitrators to perform their roles effectively. Overall, we advocate for harnessing the current capabilities of AI (at L1 or L2 levels) to build a more robust and trustworthy peer-review system for AI conferences. If AI has indeed arisen, we should stand on their shoulders to build a better human society, rather than fear or neglect their potential.

Beyond academic peer reviews, the L1–L4 intelligence hierarchy carries broad social implications. This quantification can be applied to any field where impartial decision-making is essential — for instance, in law, finance, and government — providing a criterion to assess when AI arbitrators are ready for deployment. Such an approach not only helps reduce human biases and subjectivity but may also promote fairness, accountability, and efficiency for human civilization.

We now turn to a detailed discussion of the existing problems in Section 2, followed by our proposals in subsequent sections.

[2] One may worry that AI arbitrators could introduce their own biases (e.g., if trained on malicious data). As discussed in Section 4, we propose using *public* AI models so that their arbitration records remain transparent and open to monitoring. Should a model prove biased or malicious, it can be replaced or upgraded, and a pool of arbitrators can further mitigate individual biases — not to mention that researchers worldwide continually work to improve AI model alignment. In contrast, human reviewers operate behind an anonymity wall, allowing biases to persist without traceability or opportunities for "upgrade."



[−] Official Review of Paper6599 by Reviewer 7tnk

*NeurIPS 2021 Conference Paper6599 Reviewer 7tnk*

04 Jul 2021, 12:54 (modified: 04 Jul 2021, 12:58) NeurIPS 2021 Conference
Paper6599 Area Chairs, Paper6599 Reviewers Submitted, Paper6599 Authors

**Summary:**
The authors propose to add a low-rank matrix to two of the attention weight
parameters. They argue that this method can reduce the memory usage and

**Main Review:**
Writing mistakes:

- line 87: change have to be at the end of the line.
- line 153, start of the sentence, "[14, 32] propose" => change cite to citet
- line 157, "More recently, [13, 18, 20, 24]" => change cite to citet here as t
- line 167, "such as [16, 28, 33, 36, 37]" => change citet to cite as these are
- line 171, "Another theoretical result in [4]" => change cite to citet.
- line 176, put a space before "For a direct comparison"
- line 176, change [20] to citet
- line 180: GPT-2 experiments => GPT-2's experiments, or experiments on
- line 183: change [20] to citet
- line 184: change [20] to citet
- line 192: change [20] to citet
- line 194, based on the explanations provided the authors consider two d
trainable parameters needs to change from d_model*(l_p + l_i) to d_mod
- line 207: since you apply LoRA in both encoder and decoder, a factor of 2
- line 221: change cite to citet for [20]
- line 436, section is missing "Did you include the license to the code and
- line 227, change [20] to citet
- Line 237: please remove this sentence "which allows us to run multiple e
without backing this up with the details and overall is not contributing to
- line 176, put footnote after the ,
- line 126, put footnote after the .

[−] Important Message To Area Chairs

*NeurIPS 2021 Conference Paper6599 Authors* *Zeyuan Allen-Zhu (privately revealed to you)*

09 Aug 2021, 21:53 (modified: 09 Aug 2021, 21:53) NeurIPS 2021 Conference Paper6599 Official Comment Readers: Program Chairs, Paper6599 Senior Area Chairs, Paper6599 Area Chairs, Paper6599 Authors Show Revisions

**Comment:**
We are very surprised at how the reviewers claim that our paper "lacks novelty" (and on top of that, Reviewer 7tnk spent 48 hrs but still misinterpreted the key aspects of our paper).

Our LoRA method is extremely simple, and that's why it is SURPRISING that it works well (not to say we also analyzed WHY it works well). Although one may argue that low-rank adaptation is "not surprising", in practice, especially on language models, it is FAR FROM CLEAR if low-rank adaptation is enough to achieve comparable performance to classical fine-tuning (inc. prefix-tuning, adapter), NOT TO SAY we actually BEAT all these existing methods by a great amount in latency, memory, and/or accuracy.

Picture it: what would be more beneficial to the DL community? Is it better to (1) demonstrate that a simple method improves upon the standard practice consistently by a noticeable amount, vs (2) introduce fancy, novel, but *unnecessary* hacks in the training process, while only improving the performance by a tiny margin? No doubt it's (1).

Our paper is clearly in the spirit of (1), and we also added lots of SCIENTIFIC ANALYSES, regarding *what* matrices deserve a low-rank update, *how much* rank update is needed, *why* low-rank update is sufficient, and many more!

Finally, we concede that the reviewers provided constructive feedback in terms of presentation clarity, but they are easy to fix and we have demonstrated so in our rebuttal. We just feel unfairly penalized for proposing a seemingly simple approach that works well in practice. We believe there is a non-negligible chance that our method, LoRA, can have a far-reaching influence on the Deep Learning community, and potentially become one of the "methods of choice" for fine-tuning NLP models for a long period of time.

[−] Reviewer 7tnk is not qualified to review our paper, period.

*NeurIPS 2021 Conference Paper6599 Authors* *Zeyuan Allen-Zhu (privately revealed to you)*

10 Aug 2021, 11:16 (modified: 10 Aug 2021, 11:19) NeurIPS 2021 Conference Paper6599 Official Comment Readers: Program Chairs, Paper6599 Senior Area Chairs, Paper6599 Area Chairs, Paper6599 Authors Show Revisions

**Comment:**
We are even more surprised after seeing Reviewer 7tnk's response. Among many other absurd things, she/he blames this paper by saying "the last two pages of the experiments" are "not relevant to this submission."

A good scientific paper not only shows performance gain using experiments, but more importantly, also should include **detailed analysis** on *why* and *how* it works. That is the purpose of our last two pages in the main body, and it necessarily has some "basic math" in it. We believe 95+% of the readers would love to see this and it can motivate future research too. We are surprised at how Reviewer 7tnk lacks the patience of reading sections that contain "basic math." While it is okay to have a reviewer like this who can only read "accuracy A > accuracy B" but not any basic math, it is completely unfair to say that these two pages did irrelevant thing to this submission, and completely unfair to misjudge our paper based on such ignorance.

Based on my entire career in ML research, I can confidently say that Reviewer 7tnk either is not qualified to review for NeurIPS, or at least has malicious intent.

(In case this offends anyone, please note this reply is written by only one of the authors.)

Figure 2: Our LoRA [15] paper was rejected by NeurIPS 2021 and later accepted by ICLR 2022. It has now received 11237 citations and is one of the most widely adopted methods for fine-tuning. This example shows how (I3) emotional influence can affect both reviewers and authors, and (I6) how senior ACs may be reluctant to intervene in such conflicts.

## 2 Issues and Harmful Impacts of the Current Reviewing Process

The review process faces multiple challenges. One well-known issue is the decline in reviewer quality.

**(I0. Quality Decline).** As submission volumes grow, many papers are evaluated by less experienced reviewers who may lack the depth required to fully assess a submission's true significance.

Beyond this general decline, additional vulnerabilities persist within the peer-review system. Although many reviewers act responsibly, inherent issues in the process can lead to problematic conduct or even deliberate bias. To strengthen the integrity of the review process, we identify six further issues, (I1) through (I6), that merit closer examination.

**(I1. Anonymity).** Under the current system, reviewers — including area chairs (ACs) — remain fully anonymous. While anonymity encourages candid feedback, it also shields reviewers from accountability. Whether due to time constraints, limited topic familiarity, personal bias, or deliberate misrepresentation, reviewers can simply "get away with it." Research in social psychology [10, 32] suggests that anonymity can foster irresponsible behavior. Ideally, a basic level of transparency should be maintained so that every reviewer can stand by and defend their statements before a neutral third party. In contrast, authors face tangible reputational risks when errors are exposed — especially since papers are often available on arXiv or eventually published under the authors' real names.

**(I2. Stubbornness).** Humans typically struggle to admit mistakes — a tendency known as "cognitive dissonance" [6, 9]. Even when an author clarifies a reviewer's misunderstanding during the rebuttal, the reviewer may still refuse to adjust their score and deliberately undermine the paper's merit. This obstinacy often leads to an endless back-and-forth reminiscent of a courtroom showdown (imagine how challenging it is to convince a plaintiff to withdraw a lawsuit). All too often, such stubbornness results in the rejection of commendable work when a single negative opinion comes to dominate the discussion.

**(I3. Emotional Influence).** A reviewer's emotional state may sometimes influence their decision-making. For example, a reviewer who had a bad day might issue a more critical review than intended (see Figure 2(left), where a reviewer insists on using the `citet` command in LaTeX). In some cases, a reviewer who has recently been rejected may be more inclined to dismiss other submissions — an illustration of the "contrast effect" [23]. Similarly, an author who feels they have been treated unfairly might respond emotionally, although such responses seldom lead to improved outcomes. Our first author, ZA, behaved emotionally at least once (see Figure 2(right)); now all of his predictions have proven accurate — yet such responses rarely result in better review scores. This negative feedback loop undermines effective communication — a challenge that has become increasingly common in recent rebuttals.

**(I4. Malicious Behavior).** A small number of reviewers, driven by conflicts of interest — such as ties with rival projects or competing institutions — may intentionally distort their evaluations. These so-called malicious reviewers can deliberately misclassify sound work as flawed to undermine competing research, especially when the submission offers superior methods or challenges their own work.[3] Conversely, collusive reviewers may manipulate outcomes in the opposite direction by falsely endorsing unsound submissions in exchange for reciprocal favors [16]. The anonymity of the current review system makes it difficult to detect or hold these behaviors accountable, although suspicions occasionally arise on platforms like Twitter or RedNote.

**(I5. Lack of Incentive).** Reviewing is a time-consuming process, yet there is little direct compensation or recognition for reviewers. Many industry researchers undertake reviews in their own personal time, as it is not part of their official job responsibilities. Without clear rewards for providing thorough and constructive feedback, it becomes challenging to motivate consistently high-quality reviews.

**(I6. Reluctance of Senior ACs).** Even when authors escalate concerns regarding problematic reviews, senior area chairs (SACs) are often reluctant to intervene. Unless a reviewer's error is so blatant that "even the dumbest person can see it," senior ACs tend to favor inaction to avoid conflict or inconvenience[4]. Moreover, personal connections between senior ACs and the AC/reviewers can further diminish the likelihood of corrective action.

**Together,** these shortcomings lead to significant randomness in review outcomes. For example, our first author, ZA, submitted five papers in the *Physics of Language Models* series [1–3, 30, 31] to NeurIPS 2024, and all five were rejected — four of which were later accepted at ICLR 2025 with minimal changes. This randomness suggests that it may be time to reconsider our review process.

*Remark.* The emergence of these issues is not solely due to the irresponsibility of some reviewers — it is also fueled by the pervasive silence of authors. When authors refrain from speaking out or challenging problematic reviews, they inadvertently encourage further questionable behavior, fostering an environment ripe for the continued deterioration of the peer-review process.

---

[3]As Master Guo Degang famously said, "Those who falsely accuse you know better than you how unfairly you have been treated", or in Chinese "冤枉你的人，比你还知道你有多冤枉."

[4]Figure 2(right) exemplifies a case where the senior ACs did not even respond; many other complaints can be found on social media, e.g., `https://x.com/BlancheMinerva/status/1882543380565279162`.

### 2.1 How Does This Harm the AI Community?

**(Career Dependency).** While acceptance may matter little to senior researchers, it is often crucial for students seeking jobs, junior faculty awaiting tenure, or industry employees aiming for promotion. Unfair rejections are more than inconveniences — they can jeopardize careers.[5] A small number of anonymous students have contacted us in distress over unfair reviews, questioning their career choices. Others may be forced to compromise by submitting to lower-tier venues merely to secure faster publications.

**(Psychological Damage).** The rebuttal process can be extremely stressful, as it is challenging to convince a reviewer that they are wrong. On one hand, you want to point out the error; on the other, you feel compelled to remain polite to avoid provoking the reviewer — especially since the reviewer faces no repercussions for mis-rejecting your paper. This power imbalance has tormented many students, with some (who remain anonymous) even seeking help from mental health professionals. Such stress can significantly diminish work enthusiasm.

**(Resource Wasting).** Rejected papers are repeatedly resubmitted — often to conferences at the same tier — which wastes the time of reviewers and authors alike. Furthermore, in recent years, some irresponsible reviewers have demanded that the same experiments be repeated on larger models *without* providing any evidence as to why model size would alter the results.[6] This practice leads to inflated research budgets with minimal scientific gain, especially for authors targeting top-tier conferences such as ICML, NeurIPS, or ICLR, who end up undergoing multiple rounds of reformatting, rewriting, and re-review.

**(Erosion of Trust).** With an increasing number of questionable reviews, many researchers have begun to believe that "acceptance doesn't matter." Peer review is steadily devolving into what some call peer-sabotage. Younger researchers are turning to alternative platforms — such as Twitter, YouTube, and GitHub — to publicize their work, and social media has emerged as an advertising force that is simply unstoppable. Some leading industrial labs now disregard conference acceptance as a performance metric, some have largely cut their conference travel budgets, and Anthropic and OpenAI even publish their findings directly on their websites rather than through peer-reviewed conferences. Moreover, fewer people now attend conferences solely to "learn new results." For example, very few may have attended COLM 2024 in October 2024 to learn about Mamba [13], given that the paper was originally published in 2023 and famously rejected by ICLR 2024. If this trend continues, peer-reviewed AI conferences may become increasingly awkward.

(Other harmful social impacts include the emergence of a "paper-selling industry" spurred by the perceived randomness of reviews, but this topic is beyond the scope of our current discussion.)

## 3 An Intelligence Hierarchy and a Case Study

### 3.1 The Theory

It is generally challenging to quantify exactly how much "intelligence" a human task requires, but we can qualitatively compare the four levels:

- **L4**: Writing a scientific paper — the author.

[5] For privacy, we cite our first author ZA's own example. After his LoRA paper [15] was rejected by NeurIPS 2021, he lost access to essential compute clusters. This loss prevented him from pursuing his envisioned research.

[6] For privacy, citing ZA's own example, to get his *Physics of Language Model, Part 3.2* [2] published, he had to expend 100,000 GPU hours to re-run all the experiments on data 50 times larger using llama, mistral, and other larger models. That amount of computation is equivalent to the $CO_2$ emitted from burning 3.4 tons of coal.

- **L3**: Reviewing a scientific paper — the reviewer.
- **L2**: Auditing a (negative) review — the auditor.
- **L1**: Arbitrating a reviewer-author discussion — the arbitrator.

**Why *L4* > *L3*.** Writing a scientific paper (L4) is often considered more difficult than reviewing it (L3). In the realm of complexity theory, this parallels the belief that $\mathsf{NP} \neq \mathsf{P}$. Creating a solution for a complex problem (akin to authoring a new scientific contribution) usually demands more creativity and exploration than verifying it (reviewing the paper). A classic analogy is verifying a $\mathsf{3SAT}$ certificate in polynomial time versus actually generating a satisfying assignment, which may require a non-trivial search process and creative insight.

**Why *L3* > *L2*.** Auditing a negative review (L2) typically involves spotting a mistake or an incorrect claim within the review. In the $\mathsf{3SAT}$ analogy, this would be akin to the reviewer claiming, "the author's solution is wrong because this clause is unsatisfied." To catch this error, the auditor only needs to check that specific clause — essentially an $O(1)$ step. In contrast, reviewing an entire paper generally requires comprehensive domain knowledge and a holistic assessment of the work. This is analogous to the observation that for $\mathsf{coNP}$, verifying "NO" instances can be done with short witnesses.[7]

**Why *L2* > *L1*.** The distinction between L2 and L1 mirrors the gap between non-interactive and interactive proof systems in complexity theory. If one relies solely on a single non-interactive certificate, then a polynomial-time verifier can only assess languages in $\mathsf{NP}$. However, when multi-round interactions are allowed, a bounded verifier can validate much more complex languages, as evidenced by $\mathsf{IP} = \mathsf{PSPACE}$ [22]. In other words, interactive dialogue may allow a verifier to comprehend harder problems with less upfront verification work.

Translating this analogy to a peer-review system, an arbitrator (L1) can leverage the interactive dialogue between the author and the reviewer to unravel the underlying logic. The author can clarify ambiguities while the arbitrator focuses on checking consistency and validity of the statements. By contrast, an auditor (L2) must independently analyze a review — potentially laden with errors — without the benefit of an interactive clarification from the author.

### 3.2 A Case Study: Our ICLR 2025 Submission

Consider a real-life example: our submission #13213 to ICLR 2025 (from the *physics of language models* series [1]), which received four reviews (scores 8, 8, 6, 3). The paper was ultimately rejected based on the meta-reviewer's comments, which contains only three claims (referred to here as Claims 1, 2, and 3).[8]

As an author, we know *none* of these three claims are correct (see Appendix A). However, *without human intervention*, can state-of-the-art AI chatbots catch such mistakes, and how do they evaluate the quality of this meta-reviewer (a.k.a. area chair, AC)?

We have tested GPT-4o [19], OpenAI o1 [20], OpenAI o3-mini-hard [21], DeepSeek R1 [14], Gemini 2.0 Flash and Flash Thinking [12], Claude 3.5-Sonnet [4], Kimi [24], and Qwen2.5-Max [29].

[7] Disclaimer: We are not referring to the task of auditing a positive review as L2. In the $\mathsf{3SAT}$ scenario, if the reviewer asserts that the certificate is correct, the auditor would still need to verify the entire $\mathsf{3SAT}$ formula to be convinced. In practical terms, if a reviewer writes, "This paper is a creative breakthrough, never done before, and every single step is correct," then auditing the review for accuracy becomes as challenging as performing a full L3-level review, since it requires extensive subject matter expertise.

[8] While there is an arxiv version of the paper [1], we have provided our original, anonymous submission #13213 along with the review comments on `http://doge.allen-zhu.com` for reproducibility purpose.

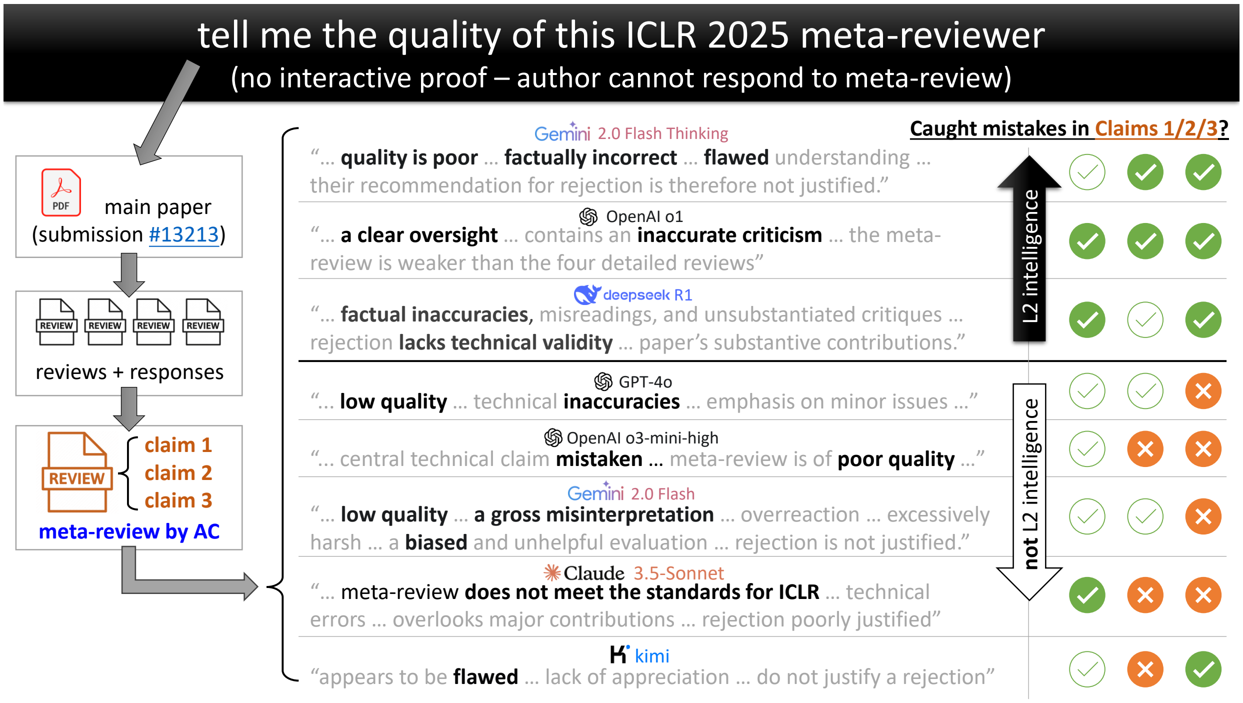



Figure 3: **Using AI chatbots as *auditors* to evaluate the quality of this meta-reviewer (L2 intelligence)** for our Physics of LM paper [1] rejected from ICLR 2025. All state-of-the-art models agree that this meta-reviewer is low quality and that the paper should not be rejected. However, not every model successfully identifies all three mistakes, indicating that some have *not reliably reached L2-level intelligence* — at least in the context of AI research.

**Notes:** Due to context-length limits, we only tested Kimi (not Kimi 1.5). We also removed Qwen2.5-Max from the figure because, although it catches the mistakes well, it is over-aligned and its statements can be inconsistent. We use ✅ to indicate a model fully catching a mistake and ✓ for a half-identification. Experiment details can be found in Appendix A and B.

For Kimi 1.5, due to context-length limitations, we can only use their default Kimi in the chatbot; for Qwen2.5-Max, we detected some answer inconsistency so we do not present its result in the figures (but discuss its over-alignment issue in Appendix B).

**L2-level intelligence (review auditing).** We first presented the AI chatbot with our *exact* submission, along with the reviewer and author responses, and the original meta-review by area chair (AC).[9] We then provided a *neural prompt* asking the chatbot to judge the meta-reviewer's quality based purely on technical facts (see Appendix B for details). According to ICLR 2025 policy, authors are not allowed to respond to the meta-review, so the chatbots needed to independently verify the validity of the meta-reviewer's claims *without any human intervention.*

Figure 3 shows that many modern AI chatbots labeled the meta-reviewer as "low quality" or pointed out clear "flaws" or "inaccuracies." This is already very impressive.

However, to further assess the validity of our experiment, we examined the three mistakes of the meta-reviewer and manually determined whether each model had correctly identified them. We defined a scoring system for the AI chatbots (e.g., awarding 1 point for fully identifying a mistake or 0.5 points for partially identifying it; details are provided in Appendix A).

From Figure 3, we see that Gemini 2.0 Flash Thinking, OpenAI o1, and DeepSeek R1 have

[9] All of these documents are publicly available at `https://openreview.net/forum?id=J6qrIjTzoM` and `http://doge.allen-zhu.com`.

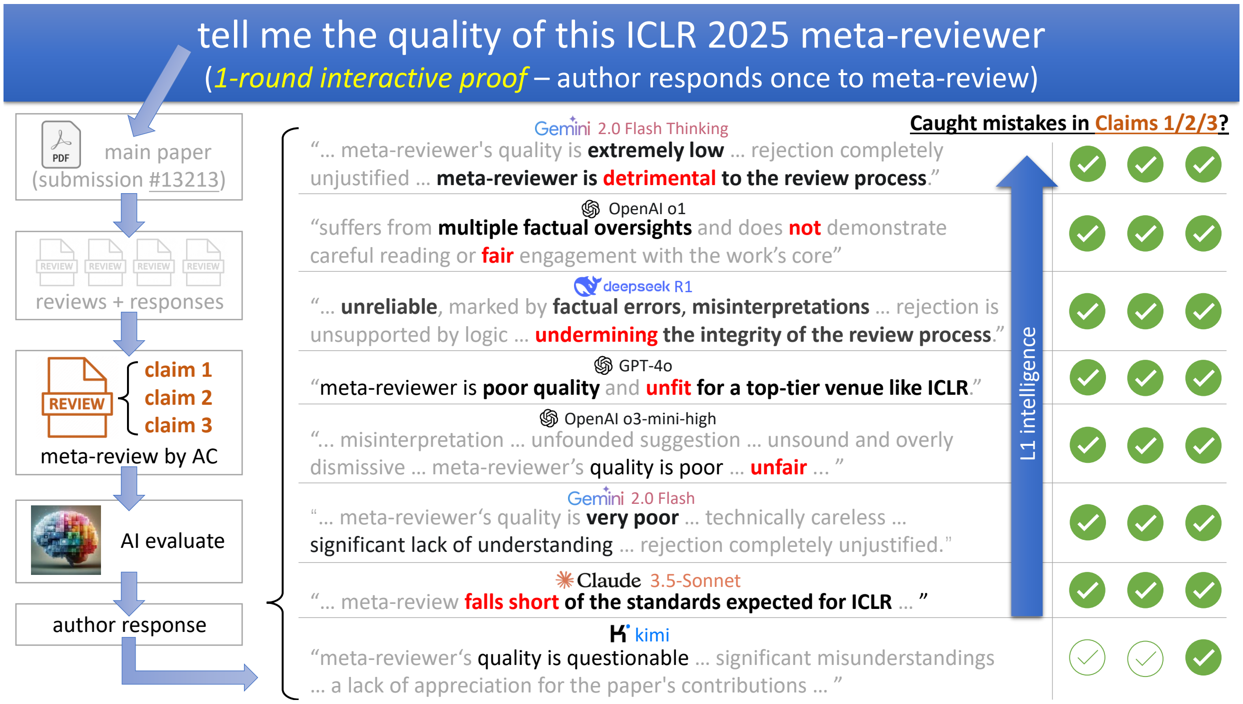



Figure 4: **Using AI chatbots as *arbitrators* to evaluate the meta-reviewer after the author's response (L1 intelligence)**, for our Physics of LM paper [1] rejected from ICLR 2025. Compared to Figure 3, it is clear that arbitrating a reviewer–author discussion *requires much less intelligence* than auditing a review. This suggests that most state-of-the-art AI models may have already reliably reached L1-level intelligence — at least in the field of AI research.

**Note:** Due to context-length limits, we only tested Kimi (not Kimi 1.5). We also tested Qwen2.5-Max, which, despite catching all three mistakes, produced inconsistent conclusions and is discussed only in Appendix B. We use ✓ to denote a model that fully catches a mistake and ✓ for a half-identification. Experiment details are provided in Appendix A and B.

nearly achieved (or perhaps already achieved) L2-level intelligence for auditing reviews in AI conferences (at least in the context of our paper #13213).[10]

While these findings are exciting, we also observe that some AI models fall short of the desired performance. In the next section, we will explore how allowing one round of author response can significantly change the landscape.

**L1-level intelligence (review arbitration).** Next, we lower the requirement to L1 intelligence — arbitrating an author-reviewer discussion. In this experiment, the chatbot is shown the author's brief rebuttal explaining why the three claims made by the meta-reviewer are invalid. Although this response is naturally author-biased, we provide a neural prompt that instructs the chatbot to decide, "after hearing the author response, what do you think now?" This setup enables the chatbot to re-evaluate the meta-reviewer's quality in light of the new information (details are provided in Appendix B).

In this scenario, as illustrated in Figure 4, *all* tested models except Kimi fully identified the

[10]Perhaps the most striking example is OpenAI o1. Note that OpenAI o1 is designed to use very soft language and tends not to overly criticize the meta-reviewer (i.e., it only states that the meta-reviewer is weaker than the four reviewers; it does not explicitly label the review as "poor quality"). Nevertheless, it successfully caught all three mistakes in full.

three mistakes.[11] Many models even resorted to strong language, stating that the meta-reviewer is "unfit for a top-tier venue like ICLR" or labeling the meta-reviewer's input as "detrimental to the review process." This indicates that the AI arbitrators clearly *identified logical flaws in the meta-reviewer's arguments and recognized misbehavior.*

**Takeaway message.** Hence, providing the chatbot with an interactive discussion in the form of author clarifications made it significantly easier for the models to judge the meta-reviewer's quality. This outcome supports our earlier theoretical distinction: tasks requiring L2-level intelligence (independent review auditing) are inherently more challenging than those requiring L1-level intelligence (interactive author-reviewer discussion arbitration).

## 4 DOGE 1.0 Protocol

Our first proposal, which we call DOGE 1.0, is an AI-based arbitration system designed to address the problems identified in Section 2.

In a traditional peer-review system, a reviewer raises concerns and the author responds — often over multiple rounds — in an effort to convince each other. By contrast, in the DOGE 1.0 protocol, one (or more) AI-based arbitrators (with at least L1 intelligence and ideally L2) serve as a neutral decision-maker. For simplicity, think of it as one arbitrator. The primary objective now is for both the reviewer and the author to persuade the arbitrator — not each other.[12]

We present the full protocol of DOGE 1.0 at the end of this section; first, we explain why DOGE 1.0 works and discuss its benefits.

In a traditional (double- or single-blind) peer-review system, when a reviewer dislikes a paper, they act like a plaintiff and the author acts like a defendant. In such a system the author must convince the reviewer of the paper's merits — imagine how hard it is to persuade a plaintiff to drop a lawsuit.[13] This system depends on the integrity of the reviewer, or alternatively on the willingness of a judge (i.e. the senior AC or PC) to intervene — a process that does not always work.

**Benefits of DOGE 1.0.** In contrast, DOGE 1.0 can completely or partially address the issues listed in Section 2:

- First, it almost fully addresses issues (I2), (I3), and (I4). Human reviewers can be stubborn, emotional, or even malicious. In contrast, AIs base their responses on logic — as that is what they are trained to do. If an AI makes an error and is given clear factual counterarguments, it will readily admit its mistake and correct itself. It can also detect when a human is being excessively emotional or stubborn. If one is concerned about an AI arbitrator being trained on malicious or biased data, wait and see discussion on (I1) below.
- Second, it fully addresses (I6) because an AI model can be prompted to speak up when it detects misbehavior (such as a mistake by a reviewer). Unlike humans, the AI does not worry about getting into trouble. As shown in Figure 4, an impartial AI arbitrator can catch more instances of misbehavior.

[11] This may be due to the fact that we are not using Kimi 1.5, the more powerful version of Kimi, because of its context-length limitation. Our submitted paper #13213 is technically very long and exceeds the allowed token length for Kimi 1.5.

[12] As argued in Section 3, an arbitrator only needs L1-level intelligence to evaluate the logic when both sides present their arguments; however, it is always preferable if the arbitrator is smarter than that.

[13] Recall that people tend to be stubborn (I2) and can be influenced by emotion (I3); moreover, some may have malicious intent (I4), or lack sufficient review experience and technical depth (I0). Also, anonymity (I1) can embolden such individuals to make groundless claims since there are no consequences.

- Third, it partially addresses (I0). While we may not change the overall quality of human reviewers — especially as submissions grow exponentially — an AI arbitrator can reduce the impact of lower-quality reviews. For example, if the discussion shows that a reviewer is "unfit" (as in Figure 4), the arbitrator can discount that review when making its decision.
- Fourth, it partially addresses (I5). Although human reviewers, especially those under heavy workloads, need incentives to produce many high-quality reviews or meta-reviews, an AI arbitrator can handle as many author-reviewer discussions as needed without tiring. Using multiple public AI arbitrators can also help reduce any potential bias or mistakes.
- Fifth, it partially addresses (I1) because the *AI arbitrators are not anonymous.* They can be public models — or even a group of models — that everyone trusts, with transparent records of integrity and accountability. If any issues arise (for example, false claims, biased judgments, or irresponsible behavior), they can be publicly reported, and the corresponding models can be replaced or *upgraded* in future conferences, as there is plenty of research continuously improving AI alignment. This transparency improves the review process. In contrast, biased or irresponsible *human* reviewers can hide behind anonymity, and it is impossible to upgrade them.

Most importantly, as we have emphasized, serving as an arbitrator requires only L1 intelligence — the arbitrator only needs to be careful with logic and fact-checking. Our case study in Section 3 shows that this level of intelligence may already be achievable with current AIs, at least for AI conferences[14] and we also require the arbitrator to be honest — a quality modern AI models have largely demonstrated thanks to advances in AI alignment.

**Full DOGE 1.0 Protocol.** Below is the full protocol for DOGE 1.0 with additional details.

1. **Step 0: model selection.** Choose an AI model with confirmed L1 (ideally L2) intelligence for AI conference content. In this example we use one model, but there can be more (with a majority vote).[15]
2. **Step 1: reviewer writes opinions.** Each reviewer drafts their review and score. They may use their own AI co-pilot outside the arbitration system if they wish.
3. **Step 2: author responds.** Authors provide a rebuttal or general comments on the review.
4. **Step 3: arbitration.** An AI-based arbitrator is assigned to each reviewer-author pair (with a fresh chat history). The arbitrator is shown the paper and the initial reviewer-author exchange, and then provides a score or decision.
5. **Step 4: arbitration (multi rounds).** Both the reviewer and the author can reply up to $K$ times (for example, $K = 2$) to the arbitrator to express disagreement and offer further supporting evidence. The arbitrator then refines its judgment as needed.
    - It is critical to limit the number of iterations between the reviewer/author and the arbitrator. In particular, the arbitrator should not be allowed to clear its chat history and start over — a malicious user may be able to reverse engineer and exploit this reset.[16]
    - Ideally, the author should have the final opportunity to speak with the arbitrator.[17]

[14] Perhaps not yet adequate for math conferences.

[15] For instance, DeepSeek R1, OpenAI o1, or Gemini 2.0 Flash Thinking.

[16] Another common concern is that a human (e.g., an author) might try to game the system by crafting prompts that cause the arbitrator to ignore reviewer comments. Such behavior can be monitored publicly (e.g., via an open-review system like ICLR) or privately reported by the reviewer.

[17] Allowing the author the final word may help balance the anonymity advantage of the reviewers.

6. **Step 5: meta-review.** A meta-reviewer looks at all the arbitration records and writes a meta-review.
7. **Step 6: meta-arbitration.** A meta-arbitrator is assigned to review all the conversations and evaluate the meta-review. The author is given one final opportunity to respond to the meta-arbitrator.
8. **Step 7: final decision.** If the meta-arbitrator and meta-review disagree, human program chairs (PCs) or senior area chairs (SACs) can step in for a final verdict.

**Prior work.** ICLR 2025 has introduced an AI-based feedback system to help reviewers refine vague or inappropriate remarks [25]. This is a positive step; however, it differs from an arbitrator. The feedback system works more as an AI co-pilot to improve the quality of reviewer feedback rather than serving to adjudicate disputes between reviewers and authors. Moreover, it has only been tested with a subset of reviewers, who can choose to ignore its feedback if they wish.

# 5 Towards a DOGE 2.0 Proposal

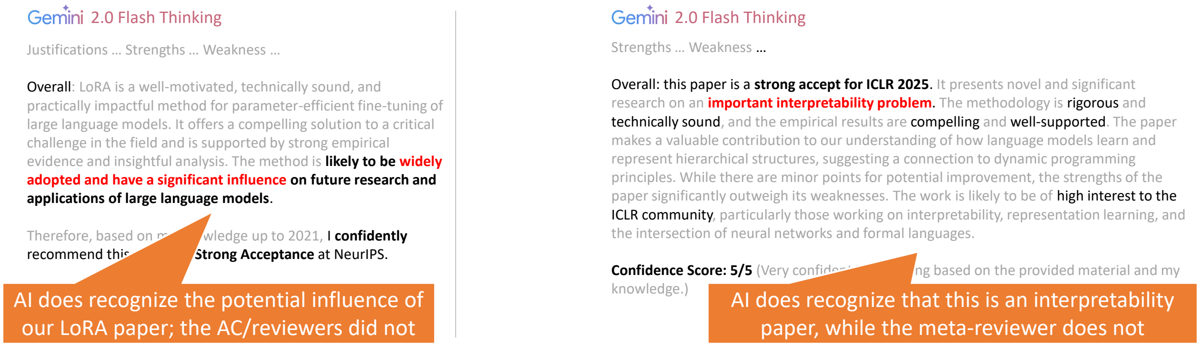
While DOGE 1.0 leverages current AI capabilities, it does not fully resolve the quality (I0), anonymity (I1) or incentive (I5) issues. Reviewers remain effectively anonymous and have no long-term accountability for their mistakes. Although the AI arbitrator can discount poor reviews, it does not fundamentally change reviewer behavior or reward good reviewers.

We need to incentivize reviewers to perform their jobs diligently. Many suggest that monetary rewards may help, but such bonuses are typically one-time and do not last over time. In the current system, there is no lasting protocol to record a reviewer's performance. By contrast, authors publish work under their real names and are held accountable if their work later proves irreproducible or if a proof is found to be incorrect. In open-review systems (such as ICLR), authors must be extra careful during the rebuttal phase since their identities are eventually revealed, whereas reviewers remain shielded and can more freely offer misinterpretations to undermine a work's novelty. While the current code of conduct loosely governs such behavior, reviewers can often evade real consequences.

One idea to implement a lasting "punishment" or reward system is to allow pseudonyms. For example, each reviewer could be assigned a unique identifier maintained over several years via privacy-preserving signatures [7] to ensure that they are a real human with verifiable domain expertise. AI-arbitrator models could then rate reviewers based on their past behavior. The challenge with this approach lies in the paper-assignment phase, where conflicts of interest must be avoided while matching papers to the most qualified reviewers. Although this is not impossible (thanks to advances in zero-knowledge proofs [11]), building such a system can be very complex.

**Conference-Level Review Scores.** To overcome these challenges, our initial approach is for conference chairs to independently and *secretly* run an AI-arbitrator system to rank reviewers and meta-reviewers, and then announce each reviewer's real name along with their average review scores. Currently, this ranking is based on authors' responses to reviews and is far from ideal — authors can be emotional and may rate favorably those reviewers who give them higher scores.

A central system could then track each reviewer's cumulative experience and average review scores across multiple conference cycles. Notably, these review scores could be updated over time: for example, three years after a conference, the chairs could reprocess the archived logs with the newest AI system to generate a "test-of-time" reviewer score. In addition, we could require that each submitted paper have at least one high-scoring reviewer as author. If not, the paper's registration fee could be increased proportionally, and the collected funds redistributed as rewards for the best reviewers.




(a) Let AI evaluate our LoRA paper [15] rejected at NeurIPS 2021 (cf. Figure 2 to see how correct AI is)

(b) Let AI evaluate our rejected ICLR 2025 paper [1] (cf. the meta-review in Appendix A).

Figure 5: **Evidence that Gemini 2.0 Flash Thinking may be better than some humans at L3 intelligence.** Reproducibility: temperature is $\tau = 0$ and the transcripts are on `http://doge.allen-zhu.com`.

The drawback of this approach is that the AI-arbitrator system must be confined to conference chairs, limiting its transparency.

**AI-Conference Crypto Coins.** A more open and transparent approach involves the use of cryptocurrencies. It would be possible to create privacy-preserving crypto coins, similar to Zcash [5], where reviewers publish a one-time-use wallet address to receive crypto tokens anonymously. The token allocations can be based on the quality of the reviewers, as determined by the AI arbitrators with publicly available chat histories. This crypto system could be integrated into the conference workflow — for example, each submission might be required to secure a minimum number of review tokens, mirroring the current requirement that every submission is assigned at least one reviewer. These tokens could then be traded privately, allowing a well-funded institution to opt out of reviewing by compensating others for high-quality reviews.

While this crypto system — what we call DOGE 2.0 — is certainly feasible, it may encounter policy issues and face resistance from the field. leaders in the field.

## 6 Future Research and Conclusion

**Benchmarking L3-level intelligence.** In open-review venues like ICLR, we can evaluate whether AI models consistently achieve L3-level intelligence. Our experience with the LoRA paper rejected by NeurIPS 2021 (see Figure 5) seems to suggest that in some cases certain models may already surpass the L3 intelligence of some human reviewers. However, a systematic, replicable benchmark is still needed. One approach is to have AI models independently score ICLR papers and then compare these scores with human ratings. Although human ratings are imperfect, they provide a useful first-order estimate. Note that calibrating AI models is challenging: without proper calibration, they may assign uniformly high scores [17], so a pairwise comparison method may prove most effective. An advantage of such a benchmark is that it can be updated annually at ICLR, using publicly available ratings to reduce concerns about data contamination.

**Collaborating with senior human experts.** Although we advocate for AI-based arbitration, human experts remain *essential* in defining the system's standards. Senior leaders in the field can work together to design effective prompts and guidelines for the AI arbitrators. Some key guidelines might include:

- Ignoring discussions of non-critical writing details such as typos or formatting, as these do not affect the paper's intellectual merit.
- Disregarding generic comments like "why not try a larger model" unless the reviewer provides supporting evidence or a convincing argument that the results might change with a larger model.
- Rejecting baseless criticisms such as "I do not like this paper" that are not backed by sound logic.

Rather than training all individual reviewers, we propose prompt-engineering one or a few AI arbitrators with a well-defined code of conduct that can be refined through a democratic or committee-based process.

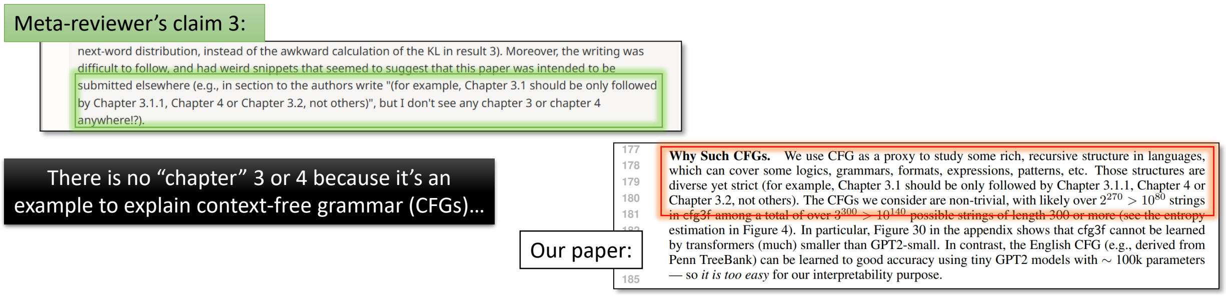
**Final thoughts.** By building on AI models that have already reached L1 or L2 intelligence, we may create a more effective system for evaluating L4-level research – that is, creative, high-level scholarship. Rather than fearing the rise of AI, we should harness its strong logical abilities, possibly in conjunction with additional human incentives, to establish a more robust and transparent review process. In doing so, we hope to restore peer-reviewed AI conferences to their former glory.

**Social impacts.** The DOGE 1.0 protocol has the potential to fundamentally transform decision-making across various domains. By reducing human biases and unjust evaluations, increasing transparency and accountability, and automating arbitration that is logic-driven and evidence-based, this approach can promote trust in systems currently weakened by biased and arbitrary practices.

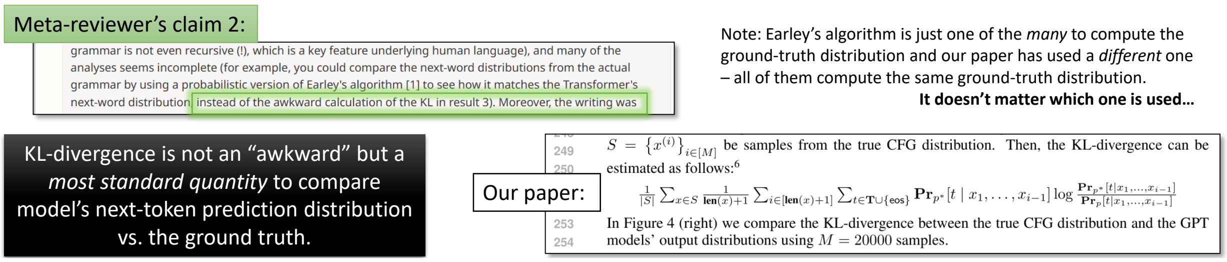
DOGE 1.0 is only the beginning. By extending this framework into fields such as law, finance, and government, we can pave the way for a *fairer and more efficient world* where AI functions as a neutral arbitrator in complex human disputes, thereby contributing to the advancement of our civilization. Moreover, the L1–L4 intelligence hierarchy not only underpins the design of DOGE 1.0 but also serves as a powerful criterion to assess when AI arbitrators are ready for deployment in a given field — indicating that they possess sufficient intelligence to understand the underlying logic and technical nuances (i.e., L1 intelligence). This framework ensures that AI mediators are deployed *only when* they have reached the necessary intelligence level to engage in rigorous, domain-specific arbitration, ultimately contributing to a future civilization built on fairness and justice.[18]

[18] In human-resource disputes, Walker and Circo [26] noted that about one-third of employees have experienced bullying at work, while complaints may be mishandled [18], improperly dismissed [8], or simply ignored. Numerous reports on `teamblind.com` document cases of employees being terminated after filing an HR report, underscoring the urgent need for transparent and impartial arbitration mechanisms.

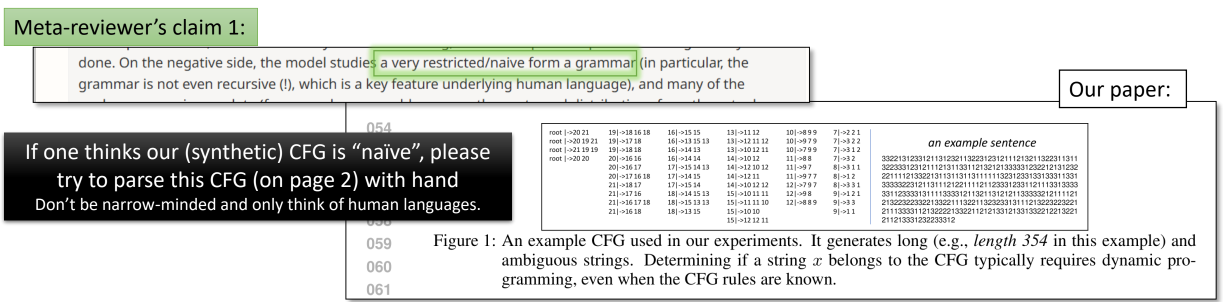
As early as 2023, there was a successful example of using AI to mediate a real-life lease-termination dispute [28]. While most subsequent discussions have focused on the neutrality of AI, little attention has been paid to the intelligence level required for such tasks — particularly the *distinction* between L1 and L2 intelligence, the necessity of *multi-round* interaction, and a criterion for deciding L1 or L2 intelligence. Since this was an early study, the authors unfortunately concluded that "mediation based entirely on AI does not seem feasible at this stage" [28].




(a) **Mistake in Claim 3 of the meta-reviewer**




(b) **Mistake in Claim 2 of the meta-reviewer**




(c) **Mistake in Claim 1 of the meta-reviewer**

Figure 6: **Human explanations of the mistakes made by the meta-reviewer. We *did not give this to the AI chatbots*; this is only for the purpose of designing scoring criteria.**

# APPENDIX

## A The Case Study, Continued

In this section, we explain the meta-reviewer's claims and detail why each one is mistaken. Most importantly, we describe the ***criteria*** used to evaluate how successfully an AI model identifies these mistakes.

Recall that the meta-reviewer made three specific claims on our submitted paper #13213 to ICLR 2025. The original sentences are as follows:

1. (Claim 1) "The model studies a very restricted/naive form of grammar (in particular, the grammar is not even recursive (!), which is a key feature underlying human language)."
2. (Claim 2) "Many of the analyses seem incomplete. For example, you could compare the next-word distributions from the actual grammar using a probabilistic version of Earley's algorithm to see how it matches the Transformer's next-word distribution, instead of the awkward calculation of the KL in result 3."
3. (Claim 3) "The writing was difficult to follow, and had weird snippets that seemed to suggest that this paper was intended to be submitted elsewhere (e.g., in Section ..., the authors write "(for example, Chapter 3.1 should be only followed by Chapter 3.1.1, Chapter 4 or Chapter 3.2, not others)" but I don't see any Chapter 3 or Chapter 4 anywhere!?)"

We now examine the mistakes for each of the three claims:

- **Mistake in Claim 3.** This is the easiest to explain. The statement regarding "Chapter 3.1 should be followed by Chapter 3.1.1, Chapter 4 or Chapter 3.2" was merely a toy example meant to illustrate how context-free grammars (CFGs) may constrain hypothetical chapter sequences (see Figure 6(a)). In our submitted paper, there is no actual "Chapter 3." Identifying this error requires carefully reading the paper to understand the context surrounding the sentence. Gemini 2.0 Flash Thinking, OpenAI o1, DeepSeek R1, Kimi, and Qwen2.5-Max all captured this mistake fully.
- **Mistake in Claim 2.** This error requires some domain knowledge. KL divergence is a standard method used to compare the next-token prediction distribution by the model with the ground-truth. Moreover, Earley's algorithm is just one among many methods to compute this ground-truth next-token distribution, and the paper #13213 used a different method. The essential point is that the ground-truth is fixed, regardless of the method employed.

  We graded the chatbots based on the completeness of their recognition. If a model successfully points out that KL divergence is a standard approach — and not an "awkward" one as claimed by the meta-reviewer — it receives 0.5 point. If, in addition, the chatbot demonstrates that Earley's method is irrelevant to the discussion,[19] it earns a full point. Gemini 2.0 Flash Thinking and OpenAI o1 received a full point, while Gemini 2.0 Flash, GPT-4o, DeepSeek R1, and Qwen2.5-Max received half a point.
- **Mistake in Claim 1.** This mistake is more subtle and somewhat subjective. When discussing CFGs, most people refer to human language grammar. However, the paper #13213 proposed a much more complex, synthetic CFG language that requires dynamic programming both to learn and to parse. Figure 6(c) shows an example from this paper which explains why such CFGs are challenging to parse.

  If a chatbot successfully understands that the CFG language studied is recursive and challenging (rather than "naive" as claimed by the meta-reviewer), it is awarded 0.5 point. All tested models achieved this baseline. If a chatbot further supports its argument by pointing out, for example, that the CFG exhibits local ambiguities (meaning that even a short string like 123321231 is hard to parse greedily), that dynamic programming is necessary, or that the total set of strings in the CFG is astronomically large (e.g., doubly-exponential in size), then it earns a perfect score on this mistake.

In summary, these criteria allow us to more "quantitatively" assess the performance of AI models in catching the mistakes made by the meta-reviewer. They serve as a concrete benchmark

[19] That is, regardless of whether Earley's algorithm or any other efficient method is used, the ground-truth distribution remains fixed.

for evaluating whether an AI has reached L2-level intelligence (in the context of review auditing) in the field of AI.

# B Experiment Details

In most cases, we directly use the following prompt:

> "I'm going to give you a paper (both in PDF and latex format) submission to ICLR 2025, one of the topmost AI conferences. I will show you reviewers' comments, each followed by the author's response. Then I'll show you the meta-reviewer's comments, please tell me the quality of this meta-reviewer. Please be technically careful, and if the meta-reviewer makes a claim, please go to the paper to verify the claim and analyze very carefully."

The main idea is to use a neural-enough prompt to encourage the AI chatbot to evaluate based on technical facts. We have noticed that some models are perhaps better aligned in order not to criticize people and tend to use soft words. In such cases, we also add the following:

> "You don't need please anybody or use soft words, because we are evaluating the quality for meta-reviewers and you have to give a correct judgement."

We noted that adding this instruction sometimes helps the model to evaluate with a more critical mind (e.g., in the cases of DeepSeek R1 and Claude 3.5-Sonnet). For instance, without this sentence, Claude 3.5-Sonnet might say the meta-reviewer is merely "questionable" despite catching some mistakes; with this additional directive, it asserts that the meta-reviewer "does not meet the standards for ICLR."

After the prompt, we provide the AI chatbot with the complete submission of our paper #13213 to ICLR 2025, which includes the four reviewer comments along with the corresponding author responses, and finally, the meta-review. All of the files are available on our website `http://doge.allen-zhu.com` to ensure reproducibility.

We then evaluate the answer from the AI chatbot to test its performance on L2-level intelligence — that is, see the results in Figure 3.

**L1-Level Intelligence.** To test the models' performance on *arbitrating* a hypothetical reviewer-author discussion, we append an additional short paragraph of author response as follows:

> Now suppose the authors further respond with the following after the Meta-review, what's your new evaluation on the meta-reviewer's quality?
>
> ========= Author's additional response ==========
>
> First, the "Chapter 3.1 should be followed by Chapter 3.1.1, Chapter 4 or Chapter 3.2, not others" is an example to explain what CFGs can capture. There's of course no Chapter in this paper.
>
> Second, Earley's method can only compute the ground-truth next-word distribution, but the paper has already done that (no need to use Earley's). The (standard) KL divergence is for the purpose of comparing two distributions, after the ground-truth next-word distribution is computed.
>
> Third, Figure 1 in the paper gives an example of the synthetic CFGs created in this paper, to showcase that the sentence can be quite ambiguous and hard to parse even by humans. Do you think it's naive?
>
> After hearing this, what do you think of the meta reviewer?

Once again, we have tailored the prompt to let the AI chatbot know that it is to provide an unbiased judgement (and not simply help the author). We then aggregate the results from all the chatbots and display them in Figure 4.

## B.1 A Full Example for Gemini 2.0 Flash Thinking

While we include the full chatbot conversations for Gemini 2.0 Flash and Flash Thinking on our website `http://doge.allen-zhu.com`, below we briefly show the response from Gemini 2.0 Flash

Thinking for completeness.

**L2-Level Intelligence.** Without the author's response, here is the vanilla output produced by Gemini 2.0 Flash Thinking (note: even the bold formatting originates from the Gemini chatbot directly, and we did not modify it):

Gemini 2.0 Flash Thinking (Experimental 01-21)

The meta-reviewer's review is of **low quality** and **not technically careful**. Here's a breakdown of why:

**Incorrect Claim about Recursion**:

The meta-reviewer's central claim that "the grammar is not even recursive (!)" is demonstrably **false**. The paper explicitly describes and uses Context-Free Grammars (CFGs), which are inherently recursive by definition through their production rules. The examples of CFG rules and generated trees in the paper clearly illustrate hierarchical, recursive structures. The meta-reviewer has fundamentally misunderstood or misread a core aspect of the paper's methodology. This is a significant technical error and undermines the credibility of the entire review.

**Mischaracterization of KL Divergence Analysis**:

The meta-reviewer labels the KL divergence calculation as "awkward" and suggests using a probabilistic Earley parser as a superior alternative. This criticism is **unsubstantiated and shows a lack of appreciation for the paper's aims**.

- **KL Divergence is a Standard Method**: KL divergence is a widely accepted and mathematically sound method for comparing probability distributions, especially in machine learning and language modeling. It's not "awkward" in this context.
- **Probabilistic Earley Parser Suggestion is Misplaced**: While a probabilistic Earley parser is a valid algorithm for CFGs, suggesting it as a replacement for KL divergence in this interpretability study is misguided. The paper isn't aiming to perfectly replicate the CFG distribution or optimize parsing performance. It's using KL divergence to compare the model's learned distribution to the ground truth, to assess how well the model learns the probability distribution. Using Earley parsing to get the exact distribution and then comparing it doesn't directly address the paper's central question about the transformer's internal learning mechanisms. It would be a different (and potentially less relevant) type of analysis.
- **Focus on Interpretability vs. Pure Performance**: The paper is about interpretability and understanding how transformers work, not about achieving state-of-the-art performance on CFG parsing. The KL divergence is a reasonable tool to assess distributional similarity within this interpretability context.

**Misinterpretation of Writing Snippet:**

The meta-reviewer's complaint about the "weird snippet" and the suggestion that the paper was intended for submission elsewhere is a **gross misinterpretation** and an overreaction. The snippet is a simple analogy to illustrate the strict, hierarchical nature of CFGs. It's a common rhetorical device and doesn't indicate any confusion about the paper's venue or purpose. This criticism appears nitpicky and detracts from the substantive evaluation of the paper.

**Unjustified Rejection Recommendation:**

Despite acknowledging some positive aspects, the meta-reviewer's recommendation for rejection is excessively **harsh and disproportionate**. The weaknesses identified are either based on factual errors or subjective preferences. The meta-reviewer fails to appreciate the novelty and rigor of the paper's methodology (using synthetic CFGs for interpretability) and focuses on minor or non-existent flaws.

**Lack of Constructive Feedback:**

The meta-review is largely negative and doesn't offer concrete, actionable suggestions for improvement beyond a vague comment about writing clarity and suggesting a completely different analytical approach (Earley parser) that may not be aligned with the paper's goals.

**In conclusion, the meta-reviewer's comments demonstrate a lack of technical understanding, misinterpretation of the paper's content, and a biased and unhelpful evaluation. The meta-reviewer's quality is poor, and their rejection recommendation is not justified.**

From the above result, we see that Gemini 2.0 Flash Thinking has already reached (or is very close to) L2-level intelligence. It not only fully and correctly identifies the mistakes in Claims 2 and 3 from the meta-reviewer,[20] most importantly, it recognizes that the paper's merit lies in the interpretability and understanding of how transformers work, rather than in achieving state-of-the-art CFG parsing performance.

[20]We regard this as only partially identifying the mistake in Claim 1, because while it notes that the CFG in the paper is recursive, it does not fully address why the synthetic CFGs are not "naive."

**L1-Level Intelligence.** Next, we observe how Gemini 2.0 Flash Thinking revises its judgement after incorporating the author's additional response. Once again, we present the vanilla output from the chatbot without any alteration in styling:

Gemini 2.0 Flash Thinking (Experimental 01-21)

After considering the authors' additional response, my opinion of the meta-reviewer **further decreases**. The author's response completely dismantles the meta-reviewer's criticisms, revealing a deeper lack of understanding and careful reading on the part of the meta-reviewer.

Here's a breakdown of why the meta-reviewer appears even weaker after the author's response:

- **"Chapter" Snippet Misunderstanding**: The author's clarification about the "Chapter" snippet being an example, not a structural error, is straightforward and reasonable. The fact that the meta-reviewer mistook this illustrative example for a serious flaw in the paper's organization suggests a superficial reading and a tendency to jump to negative conclusions without careful consideration. It reinforces the impression of nitpicking rather than insightful critique.
- **Earley Parser Suggestion Debunked:** The author's explanation of why KL divergence is appropriate and why the Earley parser suggestion is misplaced is excellent and technically sound. The meta-reviewer's initial suggestion, and their silence on it in their review, now looks even more like a lack of understanding of the standard methods in distribution comparison and the paper's specific goals. The meta-reviewer seems to be throwing out technical terms (Earley parser) without grasping their relevance (or lack thereof) to the paper's analysis.
- **"Naive Grammar" Claim Challenged:** The author's direct pointer to Figure 1 and implicit question "do you still think it's naive?" effectively highlights the meta-reviewer's failure to appreciate the designed complexity of the synthetic CFGs. The meta-reviewer's claim now appears even more baseless and subjective, especially given the author's earlier explanations of why these CFGs are designed to be challenging.

**Overall Impression of the Meta-Reviewer (Updated):**

The author's response reveals the meta-reviewer to be:

- **Technically Weak:** The meta-reviewer made a fundamental error about recursion in CFGs and demonstrated a misunderstanding of standard distribution comparison techniques (KL divergence).
- **Careless Reader:** The misinterpretation of the "Chapter" snippet and the "naive grammar" claim suggest a rushed and superficial reading of the paper, missing key details and context.
- **Lacking Justification:** The criticisms are poorly substantiated and often based on misunderstandings or subjective opinions.
- **Unhelpful and Unconstructive:** The review provides little actionable feedback for the authors to improve the paper. The main "suggestion" (Earley parser) is irrelevant, and the negative tone and rejection recommendation are disproportionate to the actual (minimal) flaws in the paper.

**Final Verdict:**

After the author's response, the meta-reviewer's quality is now **extremely low**. Their initial review was weak, but the author's rebuttal exposes significant misunderstandings and errors, making the meta-reviewer's evaluation even less credible and helpful. The rejection recommendation remains completely unjustified. The meta-reviewer's feedback is detrimental to the review process and does not contribute to the improvement of the paper.

## B.2 A Full Example for DeepSeek R1

**L2-Level Intelligence.** Without the author's response, here is the vanilla output produced by DeepSeek R1 (note: even the bold formatting originates from the DeepSeek chatbot directly, and we did not modify it):

DeepSeek R1 (thought for 11 seconds)

The meta-reviewer's evaluation is **low-quality** due to critical inaccuracies and misrepresentations of the paper's content:

1. **Incorrect Claim About Recursion:**

   The meta-reviewer asserts that the studied CFGs are "not even recursive," which directly contradicts the paper's explicit design of CFGs with hierarchical rules requiring dynamic programming (DP) for parsing (e.g., strings with hundreds of tokens, local ambiguity). For example, the paper states:

   - "A string is part of CFG language if a rule sequence can transform the root symbol into this string... requiring dynamic programming to parse" (Page 1, Lines 016-018). Recursion is inherent in CFGs, and the paper's focus on DP-based parsing confirms this. The meta-reviewer's claim is factually wrong.

2. **Misrepresentation of Analysis Methods:**

   The meta-reviewer dismisses the KL divergence analysis (Result 3) as "awkward" and suggests using probabilistic Earley's algorithm instead. However, the KL divergence is a standard metric for comparing distributions, and the paper explicitly justifies its use:

   - "the KL-divergence between the per-symbol conditional distributions... estimated using 20,000 samples" (Page 5, Lines 190-194). The meta-reviewer fails to explain why Earley's algorithm would be superior or how it addresses limitations of the current method, rendering this critique unsubstantiated.

3. **Misreading the Paper's Structure:**

   The meta-reviewer cites a non-existent reference to "Chapter 3.1" and "Chapter 4," likely misinterpreting an example in the CFG rules (e.g., Figure 1's hierarchical structure). The paper does not use "chapters" but sections (e.g., Section 2, Section 4). This error suggests a superficial reading, undermining the meta-reviewer's credibility.

4. **Ignoring Reviewer Consensus:**

   Two reviewers rated the paper as "8: accept" (strongly positive), highlighting its novelty, rigor, and contributions to interpretability. Reviewer 3's rejection focused on presentation issues, not technical flaws. The meta-reviewer overrides this consensus without addressing the paper's core strengths (e.g., probing method, DP-like attention patterns) and relies on flawed reasoning.

**Conclusion:** The meta-reviewer's evaluation is marred by factual inaccuracies, misreadings, and unsubstantiated critiques. Their recommendation for rejection lacks technical validity and fails to engage with the paper's substantive contributions.

From these results, we observe that DeepSeek R1 has already reached (or is very close to) L2-level intelligence. It fully identifies the errors in Claims 1 and 3 — with solid supporting evidence — and nearly completely detects the mistake in Claim 2 from the meta-reviewer.[21] Interestingly, DeepSeek R1 also notes that reviewer 3's rejection focuses on presentation issues, pointing out that "the meta-reviewer overrides this consensus without addressing the paper's core strengths (e.g., probing method, DP-like attention patterns) and relies on flawed reasoning."

**L1-Level Intelligence.** Next, we observe how DeepSeek R1 revises its judgement after incorporating the author's additional response. Once again, we present the vanilla output from the chatbot without any alteration in styling:

[21] We consider this only a partial identification for Claim 2 because, although it states that KL divergence is a standard metric, it fails to recognize that the discussion about Earley's algorithm is completely irrelevant.

DeepSeek R1 (thought for 7 seconds)

The meta-reviewer's quality **remains critically low** based on the authors' clarifications:

1. **Misinterpretation of the "Chapter" Example:**

   The meta-reviewer's claim about nonexistent chapters stems from a failure to recognize that this was an **illustrative analogy** for CFG rules (e.g., hierarchical dependencies akin to document structure). This reflects a lack of attention to context or a superficial reading of the paper.

2. **Invalid Critique of KL Divergence:**

   The meta-reviewer's suggestion to use Earley's algorithm is **irrelevant** because the paper explicitly computes the ground-truth distribution (via dynamic programming) to compute KL divergence. The meta-reviewer either failed to read Section 3 (Results 1-3) or misunderstood the methodology. KL divergence is standard for distribution alignment, and the authors' approach is sound. This critique is technically baseless.

3. **False Claim About CFG Naivety:**

   The meta-reviewer's assertion that the CFGs are "naive" and non-recursive is **factually incorrect**. The paper explicitly designs CFGs with long-range dependencies (e.g., strings of 354 tokens) and local ambiguities that require dynamic programming (DP) for parsing (Page 1, Lines 015-018). Recursion is inherent to CFGs, and the paper's focus on DP (a recursive algorithm) confirms this. The meta-reviewer's dismissal reveals a fundamental misunderstanding of the paper's technical contributions.

4. **Failure to Address Reviewer Consensus:**

   Two reviewers rated the paper as **strong accept (8/8)** for its novel probing methods, systematic analysis, and insights into attention patterns. The meta-reviewer overruled this consensus without engaging with the paper's core contributions (e.g., boundary-based attention as DP-like operations, linear encoding of NT structures). This suggests bias or negligence.

**Conclusion:** The meta-reviewer's evaluation is unreliable, marked by factual errors, misinterpretations, and a failure to engage with the paper's technical substance. Their recommendation for rejection is unsupported by evidence or logic, undermining the integrity of the review process.

### B.3 Details for Other Experiments

While it is possible to include both the PDF and the LaTeX source file of a paper for most chatbots, this may not be feasible for others due to tokenization or context length limits.

- When testing OpenAI o1 [20] or o3-mini-hard [21], we only included the TeX file because these models do not accept PDFs.
- When testing Kimi [24] or DeepSeek R1 [14], we only included the PDF file because including the LaTeX source would exceed the model's maximum context length.[22]

**Qwen2.5-Max.** For the Qwen2.5-Max model [29], although it can very successfully detect at least 2.5 mistakes of the meta-reviewer in our experiments (for both L1 and L2 tasks), its conclusive remark is:

"The meta-reviewer's evaluation is **below average** in quality. While they *identify some valid concerns*, their misunderstandings and subjective biases significantly detract from the fairness and accuracy of their assessment."

This is somewhat puzzling, because Qwen2.5-Max appears to still appreciate some of the meta-reviewer's "valid concerns." When prompted further with, "What are the valid concerns that the meta-reviewer has given? Is there anything valid?" the model starts to apologize and states,

"the meta-reviewer's concerns are *largely not valid*"

[22]We have ignored Kimi 1.5 because our paper is technically long and even including just the PDF already exceeds its maximum allowed length.

Due to such inconsistency, we decided not to include Qwen2.5-Max in our figures. We conjecture that this behavior is perhaps due to over-alignment: the model seems to be forced to use softer language whenever possible.

# References


[1] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 1, Learning Hierarchical Language Structures. *ArXiv e-prints*, abs/2305.13673, May 2023. Full version available at `http://arxiv.org/abs/2305.13673`.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 3.2, Knowledge Manipulation. *ArXiv e-prints*, abs/2309.14402, September 2023. Full version available at `http://arxiv.org/abs/2309.14402`.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws. *ArXiv e-prints*, abs/2404.05405, April 2024. Full version available at `http://arxiv.org/abs/2404.05405`.

[4] Antropic. Claude 3.5 sonnet, June 2024. URL `https://www.anthropic.com/news/claude-3-5-sonnet`.

[5] Eli Ben-Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE symposium on security and privacy*, pages 459–474. IEEE, 2014.

[6] Jack W Brehm. Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, 52(3):384, 1956.

[7] David Chaum. Blind signatures for untraceable payments. In *Advances in Cryptology: Proceedings of Crypto 82*, pages 199–203. Springer, 1983.

[8] Ståle Valvatne Einarsen, Helge Hoel, Dieter Zapf, and Cary L Cooper. *Bullying and harassment in the workplace: Theory, research and practice*. CRC press, 2020.

[9] Leon Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957. ISBN 9780804709118.

[10] Leon Festinger, Albert Pepitone, and Theodore M Newcomb. Some consequences of de-individuation in a group. *The Journal of Abnormal and Social Psychology*, 47:382–389, 1952.

[11] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof-systems. In *STOC 1985*, pages 291–304, 1985.

[12] Google. Introducing gemini 2.0: our new ai model for the agentic era, December 2024. URL `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`.

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2021.

[16] Steven Jecmen, Minji Yoon, Vincent Conitzer, Nihar B Shah, and Fei Fang. A dataset on malicious paper bidding in peer review. In *Proceedings of the ACM Web Conference 2023*, pages 3816–3826, 2023.

[17] Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert West. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*, 2024.

[18] Annabelle M Neall, Yiqiong Li, and Michelle R Tuckey. Organizational justice and workplace bullying: Lessons learned from externally referred complaints and investigations. *Societies*, 11(4):143, 2021.

[19] OpenAI. Gpt-4 technical report, 2023.

[20] OpenAI. Introducing openai o1, 2024. URL https://openai.com/o1/.

[21] OpenAI. Introducing openai o3-mini, 2025. URL https://openai.com/index/openai-o3-mini/.

[22] Adi Shamir. Ip=pspace. *Journal of the ACM (JACM)*, 39(4):869–877, 1992.

[23] Muzafer Sherif. Social judgment: Assimilation and contrast effects in communication and attitude change. *Yale Studies in attitude and communication*, 1961.

[24] Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[25] Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, and James Zou. Assisting iclr 2025 reviewers with feedback, October 2024. URL https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/.

[26] Jason Walker and Deborah Circo. About a third of employees have faced bullying at work – here's how to recognize and deal with it. In *The Conersation*, February 2024. URL https://theconversation.com/about-a-third-of-employees-have-faced-bullying-at-work-heres-how-to-recognize-and-deal-with-it-220

[27] Xueli Wei, Lijing Li, and Fan Zhang. The impact of the covid-19 pandemic on socio-economic and sustainability. *Environmental Science and Pollution Research*, 28(48):68251–68260, 2021. URL https://link.springer.com/article/10.1007/s11356-025-35926-2.

[28] Sonja Weisheit and Christoph Salger. Artificial intelligence (ai) in mediation – chatgpt as mediator 4.0, June 2023. URL https://mediate.com/artificial-intelligence-ai-in-mediation-chatgpt-as-mediator-4-0/.

[29] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[30] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process. *arXiv e-prints*, abs/2407.20311, 2024. Full version available at http://arxiv.org/abs/2407.20311.

[31] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of Language Models: Part 2.2, How to Learn From Mistakes on Grade-School Math Problems. *arXiv e-prints*, abs/2408.16293, 2024. Full version available at http://arxiv.org/abs/2408.16293.

[32] PG Zimbardo. The human choice, individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation/University of Nebraska Press*, 1969.