Position: Machine Learning Conferences Should Establish a "Refutations and Critiques" Track

Rylan Schaeffer* Stanford
Joshua Kazdan Stanford
Yegor Denisov-Blanch Stanford
Brando Miranda Stanford
Matthias Gerstgrasser
Susan Zhang
Andreas Haupt Stanford
Isha Gupta ETH Zürich & Stanford
Elyas Obbad Stanford
Jesse Dodge Allen Institute for AI
Jessica Zosa Forde Brown University
Francesco Orabona KAUST
Sanmi Koyejo Stanford
David Donoho Stanford

Abstract

Science progresses by iteratively advancing and correcting humanity's understanding of the world. In machine learning (ML) research, rapid advancements have led to an explosion of publications, but have also led to misleading, incorrect, flawed or perhaps even fraudulent studies being accepted and sometimes highlighted at ML conferences due to the fallibility of peer review. While such mistakes are understandable, ML conferences do not offer robust processes to help the field systematically correct when such errors are made. This position paper argues that ML conferences should establish a dedicated "Refutations and Critiques" (R&C) Track. This R&C Track would provide a high-profile, reputable platform to support vital research that critically challenges prior research, thereby fostering a dynamic self-correcting research ecosystem. We discuss key considerations including track design, review principles, potential pitfalls, and provide an illustrative example submission concerning a recent ICLR 2025 Oral. We conclude that ML conferences should create official, reputable mechanisms to help ML research self-correct.

1 Introduction

Advancement of scientific knowledge is inherently a cycle of discovery and refinement (Kuhn, 1962). Within machine learning (ML), the rapid proliferation of research combined with the fallibility of peer review occasionally results in dissemination and even commendation of research that is misleading, incorrect, flawed, or potentially fraudulent. While such outcomes are understandable, their occurrence is undesirable, and a critical gap exists in the current ML conference ecosystem: there is no official mechanism for critiquing problematic publications and rectifying the scientific record.

This absence of an official corrective mechanism causes problems: inaccuracies can propagate within the scientific record, fields can putter around under the illusion of progress, or critiques can be relegated to informal channels lacking rigorous evaluation. Minimizing such problematic publications by reforming the review process is unlikely to be a solution due to inadequate incentives for reviewers and institutional comfort with peer review as an efficient and reasonably well-trusted process.

*Correspondence to rschaef@cs.stanford.edu,sanmi@cs.stanford.edu, donoho@stanford.edu.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Position Paper Track.

In light of these challenges, this position paper champions the creation of a "Refutations and Critiques" (R&C) Track within ML conferences to provide an official mechanism for the field to self-correct. The objective is to provide the community with a high-profile, reputable platform dedicated to the critical re-examination of prior research published in ML conferences. Researchers would be encouraged to submit responses and criticisms of others' research as well as of their own research; authors of critiqued papers could play a critical role in the peer review process to ensure fairness and quality of submitted refutations and promote a structured scholarly exchange.

In this position paper, we elaborate on the critical facets of this proposal, including the reasoning for the "Refutations and Critiques" designation, the necessity of a distinct track, the suitability of ML conferences as hosting venues, and the foundational principles for its peer review rubric. We also explore potential pitfalls and advise how they can be preempted. To provide a concrete understanding of envisioned submissions, we conclude by presenting an illustrative manuscript examining a recent ICLR 2025 Oral. While we fully stand behind the analysis of the example manuscript, we specifically ask readers to evaluate it as an example of what submissions to the R&C could look like. Our hope is that this proposed track will supercharge ML research to rectify itself and continuously improve.

As an important disclaimer, certain observations in this work draw on community anecdotes and personal experience because peer-reviewed evidence on post-publication critiques is scarce. These valuable yet undocumented insights are precisely the type of scholarship our proposed R&C Track seeks to legitimize and disseminate.

2 The Case For a Corrective Mechanism at Machine Learning Conferences

We begin by establishing existing problems in the machine learning (ML) research status quo:

- The fallibility of peer review means that ML conferences sometimes accept and even highlight misleading, incorrect, flawed, or sometimes fraudulent research, causing false information to proliferate (Sec. 2.1).
- ML conferences currently lack official processes for rectifying such errors once they are part of the scientific record (Sec. 2.2).
- Reforming the peer review process to minimize or prevent problematic research is likely impractical due to reviewer incentives and institutional trust in the status quo (Sec. 2.3).
- Lack of formal recourse harms the field, imposes costs, and relegates disputes to non-scientific venues without impartial adjudication (Sec. 2.4).

2.1 Machine Learning Conferences Accept and Highlight Flawed Research

While the peer review process is central to research, nearly every machine learning (ML) researcher has stories of personal challenges they've encountered with peer review; we compile some common complaints in Appendix D. However, we are most concerned with a specific shortfall: failures in the review process lead to research that is misleading, incorrect, flawed, or perhaps even fraudulent. These papers are of special concern because when they receive acceptance and at times prestigious recognition, they impose significant costs on research. Specifically, such papers:

- incentivize others to produce less rigorous and overly hyped research,
- divert attention and resources from more rigorous work,
- waste future researchers' time and energy building on flawed foundations,
- pollute the scientific record and misdirect future work,
- generally undermine public trust in ML research.

Over the years, there have been multiple instances of top ML papers being publicly contested by reputable researchers:

- In privacy, the ICML 2022 Outstanding Paper "Privacy for Free: How does Dataset Condensation Help Privacy?" claimed that a technique called data condensation could be used to train performant and private models (Dong et al., 2022). Carlini et al. (2022) responded that statistical tests show the method does not improve the privacy over a naive baseline

and the technique is dominated by a standard and widely used baseline (Abadi et al., 2016). Moreover, Carlini et al. (2022) showed that the theoretical results are meaningless because one of the proof assumptions is stronger than the result (Feldman, 2022).

- In post-training of language models, a foundational OpenAI publication "Learning to Summarize from Human Feedback" introduced an analytically-derived equality for the KL Divergence between a probability distribution and its corresponding "Best-of-n" distribution (Stiennon et al., 2020), which allowed for studying many interesting effects such as reward model overoptimization (Hilton & Gao, 2022; Gao et al., 2023; Coste et al., 2024) and others (Bai et al., 2022). However, the equation for the KL divergence was actually a loose upper bound (Beirami et al., 2025; Mroueh, 2024). To the best of our knowledge, the original studies have not been re-examined to see how correct calculation of the KL Divergence changes the original results (if at all).
- In parameter-free optimization, the ICML 2023 Outstanding Paper "Learning-Rate-Free Learning by D-Adaptation" (Defazio & Mishchenko, 2023) exhibited a "new method" for parameter-free optimization of Lipschitz functions, and offered provable guarantees of convergence. Orabona (2023) claimed that the core contribution of the paper was established eight years prior, and stronger guarantees for similar algorithms already existed in the literature. Therefore, Orabona (2023) asserted that, when viewed in light of prior work, Defazio & Mishchenko (2023) exhibited no added novelty.
- In neuroscience-inspired AI, a Nature publication (Banino et al., 2018) and two NeurIPS Spotlights (Sorscher et al., 2019; Nayebi et al., 2021) claimed that neural representations matching those found in the mammalian brain emerge naturally in artificial neural networks trained on a task called path integration (wik, n.d.). However, Schaeffer et al. (2022, 2023b) showed that such neural representations did not emerge due to the task or other relevant constraints, but were instead inserted into the artificial networks by the researchers in subtle ways. Moreover, the way in which the neural representations were inserted contradicted key properties of the relevant biological neural circuits (Schaeffer et al., 2023a).
- Not all refutations grow out of mistakes or misconduct; some involve differences in scientific assumptions and experimental methodologies that vastly change projected outcomes. In model-data feedback loops, a Nature publication Shumailov et al. (2024) and an assortment of well-cited conference papers (Alemohammad et al., 2023; Dohmatob et al., 2024) promoted the viewpoint that future deep generative models models will swiftly and drastically degrade as they are trained on data generated by earlier generative models. However, Gerstgrasser et al. (2024); Kazdan et al. (2025); Dey & Donoho (2024) showed that these papers unrealistically trained new models primarily on data from the most recent generative model, which is tantamount to throwing away the majority of the internet and training GPT-5 on the outputs of GPT-4; when models are instead trained on a more realistic mixture of real and synthetic data, models demonstrate little-to-no degradation over time.

There are, of course, many additional examples that could be mentioned, e.g., (Maini & Suri, 2025; Rando, 2025; Chandak et al., 2025; Golechha et al., 2025; Ivanova et al., 2025). In cases like these, where a refutation is written, community members can more easily track the scientific conversation. In other cases, it is more difficult. To give a prominent example, Le & Mikolov (2014)'s "Doc2Vec" has been cited nearly 14.000 times, but the second author has publicly stated that the results by the other author were not reproducible (User, 2015); however, finding this out requires laboriously stitching together comments scattered across the internet (User, 2024, 2014) to gain a comprehensive understanding; such effort was done by Raff & Farris (2023).

At this point, one could reasonably criticize the above points for being non-academic un-scientific hearsay from the internet. Some of the above responses to existing papers are blog posts, Twitter threads, or preprints, which limits the authority of their findings due to lack of scrutiny by independent experts. At the same time, there is no clear place for such contributions in ML conferences.

2.2 Machine Learning Conferences Lack Mechanisms for Rectifying the Scientific Record

When published papers are later identified as misleading, incorrect, flawed or fraudulent, the largest ML conferences generally lack processes to address such situations. ML conferences lack editorial boards with the prerogative to remove clearly flawed papers post-acceptance, and they also lack

mechanisms for authors or trusted parties to attach corrigenda or errata. For egregiously flawed papers where retraction is appropriate, nothing can compel authors to withdraw their work, and voluntary withdrawal after acceptance is culturally abnormal in ML, even when mistakes are acknowledged.[2] Lastly, most ML conferences offer no official mechanism for reputable researchers to publicize discovered mistakes or problems alongside publications; the closest example is ICLR permitting public comments via OpenReview, but as we argue in Sec 2.4, public comments have low visibility, carry little weight, and again lack verification by independent experts. The structure of ML conferences, as well as their transience, produces these problems: unlike journals that can force retraction or post errata, ML conferences leave the decision to correct flawed work up to the authors.

2.3 Reforming Peer Review at ML Conferences to Prevent Flawed Research Is Impractical

Like most processes, peer review at ML conferences is fallible, but meaningfully overhauling it to minimize or prevent publication of flawed research faces obstacles on at least three fronts:

Inadequate Incentives to Improve Review Quality Peer review at ML conferences often lacks incentives since reviewers are uncompensated volunteers or conscripts fulfilling reciprocal reviewing requirements. To the best of our knowledge, there are only two incentives to encourage high-quality reviews, which come in the form of a carrot and a stick. The carrot is the "Best Reviewer Award", which recognizes outstanding reviewers with free conference registration. The stick is the "Responsible Reviewing Initiative," which desk-rejects authors' papers if their reciprocal reviews are low quality (Lee, 2024). This consequence is infrequently enforced, and whether it can make a meaningful difference at scale remains to be seen (Koniusz et al., 2025).

Inadequate Time for Rigorous Scrutiny The fast-paced timeline for peer review at ML conferences means that even the most meticulous and motivated reviewers would still only have several weeks to scrutinize submissions. In our experience, discovering deep flaws and conclusively demonstrating their existence and significance properly takes much longer than several weeks. A refutations track would allow invested researchers to scrutinize existing published work over a longer time-frame before publishing a critique in a future conference.

Insufficient Number of Qualified Reviewers This year, NeurIPS received over 27 000 submissions (CTOL Digital Solutions, 2025), a number that has multiplied ten times in less than ten years. The number of submissions has increased far faster than the population of qualified PhD students, professors, and industry researchers who can provide trustworthy reviews of these papers. Area chairs (ACs) have noted the growing difficulty of assigning papers to qualified reviewers (Reddit, 2024); inevitable failures to do so have led to the acceptance of more faulty research.

Lack of a Better System Despite complaints, ML conference peer review is a time-tested process, and we are unaware of more reliable alternatives for vetting large volumes of papers on a tight timeline. Consequently, we will argue that post-hoc mechanisms for rectifying the scientific record present a more practical solution than overhauling the imperfect-but-time-tested reviewing system (Sec. 3).

2.4 Lack of Mechanisms for Refuting or Correcting Flawed Publications Hinders the Field

Without appropriate channels to challenge published research, researchers are left with several suboptimal options. We enumerate each option below with corresponding anecdotes. We acknowledge that there are likely two sides to each anecdote, but without undergoing scrutiny by independent experts, where the truth lies may be unclear.

Silence $\rightarrow$ Persistence and Propagation of Errors Without an avenue to express valid criticisms, they go unvoiced or unpublished, thereby propagating incorrect knowledge that can misdirect subsequent research. For one recent example, Liu et al. (2024) proposed a novel optimizer claiming to improve over a widely-used baseline optimizer (Kingma & Ba, 2014; Loshchilov & Hutter, 2019), but the baseline was improperly tuned (Jordan, 2024). The researcher who discovered this error publicly

[2]Authors cannot be compelled to withdraw after acceptance based on the 2025 guides for reviewers, area chairs, and senior area chairs (NeurIPS Foundation, 2025a,c,b).

said that clarifying this detail via a publication would not be valued by the research community and consequently never documented the finding outside of Twitter (Jordan, 2025).

Speak Only When Harms Can No Longer Be Ignored → Damage Is Already Done If the consequences of problematic research become too overwhelming to ignore, then some may feel compelled to speak. One prominent example is Agarwal et al. (2021), which questioned whether the field of deep reinforcement learning (RL) was actually making progress. The authors evaluated over 16 high profile publications from up to 6 years prior and concluded that many deep RL algorithms were not improvements, as had been previously claimed. While the contribution was extremely valuable, the field had lost significant time and resources based on a false illusion of progress.

Formal Public Comments on OpenReview → Low Visibility and Little Weight ICLR permits researchers to comment publicly during the review process. While valuable, such public comments carry insufficient weight to formally amend the scientific record and possess inadequate visibility to alter the scientific record. For a recent example, Schaeffer (2024) demonstrated that an ICLR 2025 submission intentionally suppressed contradictory scientific evidence; the Area Chair then rejected the submission, but was overruled, and the submission was selected for a Spotlight. For another example regarding the NeurIPS 2024 Best Paper Runner-Up, see Kirsch (2024b).

Non-Traditional Publication Venues → Insufficient Visibility Researchers can submit articles to less traditional publication venues such as Distill (Distill, 2016) or to newer, relatively-infrequent and less well-known venues such as the Machine Learning Reproducibility Conference (MLRC). Though these are respected and peer-reviewed venues, they are younger and do not yet have the same visibility as the large ML conferences. As an example, Distill published a critique of the paper "Adversarial Examples are not Bugs, They are Features" (Ilyas et al., 2019) called "Adversarial Examples are Just Bugs, Too" (Nakkiran, 2019), which has just 2% the number of citations as the original paper.

Informal Public Comments on Blogs/Social Media → Digital Frontier Justice When robust mechanisms for scientific correction are absent, researchers may resort to informal channels like blogs or social media to voice refutations, including some examples shared in Sec. 2.1 as well as in Kirsch (2022). However, this "digital frontier justice" is fraught with problems that undermine genuine scientific discourse. Critiques aired on these platforms bypass scrutiny by independent experts and lack archival permanence crucial for the scientific record. Instead of fostering reasoned debate, these forums (especially social media) can become popularity contents, where debates are swayed by an individual author's personal reach rather than the scientific validity of their argument. Scientific disagreements often devolve into ad hominem attacks, creating uncertainty as to whether the criticisms are based on legitimate academic concerns or personal grievances. Additionally, these informal debates exclude researchers who are not active on these specific platforms, narrowing the scope of discussion and potentially reinforcing biases.

Such situations create a damaging double-bind for the scientific community: On one hand, unvetted accusations can unfairly tarnish the reputations of researchers who have produced valid and original scientific work. On the other, authors who publish flawed research can exploit an inherent asymmetry: their incorrect work carries the imprimatur of peer review, while the informal critiques against it do not. This allows them to simply deny or dismiss valid concerns until public attention wanes.

The consequences of this uncertainty are significant. It delays scientific progress as the community struggles to discern truth from noise. This burden falls disproportionately on early-career researchers. Lacking the extensive experience to easily differentiate robust findings from flawed ones, they may waste valuable time and resources trying to build upon an unsound foundation. Speaking from experience, the inability to reproduce work originating from renowned labs or championed by prominent scientists can lead to unfair criticism and profound self-doubt among junior researchers.

3 Proposal: Establish a "Refutations and Critiques" (R&C) Track

Rather than seeking to change peer review at ML conferences, we instead aim to leverage peer review as a familiar and reasonably trusted process to equip the field of ML with an official, reputable, and equally prestigious mechanism to address, critique and/or rectify flawed prior research.

3.1 Overview of Proposed "Refutations and Critiques" (R&C) Track

We propose the establishment of a dedicated "Refutations and Critiques" (R&C) track within ML conferences such as NeurIPS, ICML, and ICLR. The objective of this R&C Track is to provide a high-profile, reputable, and rigorously peer-reviewed platform for research that identifies, analyzes, and corrects misleading, incorrect, or potentially fraudulent claims presented in impactful ML publications. Officially integrating critical scholarship into main conferences would recognize this work as an indispensable component of the scientific process and would serve to enhance the overall integrity and reliability of contributions to the field of machine learning. Our vision for the R&C Track is to cultivate a dynamic and robustly self-correcting research ecosystem for the ML research community. This track would naturally be integrated with existing conference structures, including timelines and review processes, and would likely include a pilot to evaluate whether the proposed track is effective.

3.2 Why the Name "Refutations and Critiques?"

The choice of the name "Refutations and Critiques" over alternatives such as "Reproducibility" reflects a broader scope and a stronger stance on the nature of critical scientific discourse. While issues of reproducibility are undeniably important and would fall under the purview of this track, the title "Refutations and Critiques" is intended to encourage a more diverse level of engagement with prior work. Situations arise where the original results may be reproducible and technically correct, but the surrounding narrative may be (even unintentionally) incomplete, misleading or incorrect. Such situations have appeared across multiple topics, ranging from machine learning security (Carlini & Wagner, 2017; Athalye et al., 2018; Carlini et al., 2021, 2022; Tramèr et al., 2024) to deep reinforcement learning (Agarwal et al., 2021) to model collapse (Gerstgrasser et al., 2024; Kazdan et al., 2025; Dey & Donoho, 2024; Schaeffer et al., 2025a) to predictable scaling of language models (Schaeffer et al., 2023d, 2025b) to biological neural representations for spatial navigation (Schaeffer et al., 2022, 2023c), to meta-learning (Tian et al., 2020; Raghu et al., 2020; Miranda et al., 2022, 2023; Chen et al., 2020). To use model collapse as a recent example, many previous papers made a single critical assumption that training on more synthetic data necessarily results in training on vastly fewer real data in every training iteration. Based on that assumption, these works drew strong conclusions that future generative models are doomed. While the conclusion indeed follows from the stated assumption, the assumption itself is arguably highly unrealistic, and subsequent critical work demonstrated that by adjusting the assumption to be more realistic, the predicted collapse of future generative models is substantially mitigated or disappears entirely (Gerstgrasser et al., 2024; Kazdan et al., 2025; Dey & Donoho, 2024). Such research moves beyond a simple reproducibility check to a substantive critique of the generality of prior works' conclusions.

A related reason why "refutations and critiques" are appropriate is that if a paper lacks definitions or concrete claims, then refutations or failed reproductions have no meaning. This point was made by Carlini (2020), who challenged an ML privacy paper called InstaHide and wrote: "One of the core tenets of modern science [...] is that claims should be refutable [...] Unfortunately, InstaHide does not make falsifiable claims. [...] It defines an algorithm, and says it is private, without ever saying what that means. As a result, it's impossible to ever write a paper that claims to break it, because defining an attack necessarily requires a definition to break."

One may wonder: If prior critical work has been published, why then is a standalone track merited? We return to answer this question below in Section 3.3.

Thus, while reproducibility assessments are welcome as submissions, this R&C Track aims to construct a broader aegis to explicitly encourage submissions that seek to rectify the field. We seek to foster a space for rigorous, evidence-based challenges to prior research, whether they pertain to inadequate reproducibility, problematic assumptions, flawed experimental methodologies, incorrect analyses, misleading interpretations, or erroneous conclusions.

3.3 Why Is A Standalone Track Merited?

In the past, critical research has been submitted to and accepted at ML conferences' Main Tracks. Why is a dedicated R&C Track merited? There are several reasons:

Reduce Reviewer Indifference and Opposition Although critical responses can be accepted, e.g. Santurkar et al. (2018), and even awarded, e.g. Agarwal et al. (2021), such contributions are oftentimes harder to get through the peer review process. In many cases, reviewers or area chairs are indifferent to the prior work being criticized, and so are unlikely to find the results compelling and even less likely to find the results worthy of recognition. As one anecdote, Kirsch (2024a) submitted to TMLR (instead of ML conferences) because TMLR explicitly invites reproducibility studies. Even so, he had to push back forcefully against reviewer indifference to be accepted. In less common cases, reviewers or area chairs are affiliated with the prior work being criticized, meaning the reviewers are no longer neutral evaluators, and submitting authors can face strong opposition.

Create Bespoke Reviewing Standards Responses and critiques have unique aspects that are best served via unique standards of evaluation. For instance, additional specific checks may be required (e.g., whether the criticized paper's authors are aware of the objections and have been given proper opportunity to respond) and double-blind review may not be realistic (e.g., if public interactions such as GitHub issues are referenced for evidence).

Link Accepted Papers to Prior Work A response or a critique is more than a citation; it should be explicitly linked to the publication being critiqued, similar to how corrigenda and errata are.

For a guiding example, we can look to the NeurIPS Datasets and Benchmarks (D&B) Track (Vanschoren & Yeung, 2021). Prior to the creation of the D&B track, datasets and benchmarks were submitted to the NeurIPS Main Track; however, NeurIPS acted to carve out a unique track for reasons that echo what we have stated here. NeurIPS recognized that datasets and benchmarks play a fundamental role in ML research, and that inadequate and insufficiently rigorous incorporation of datasets and benchmarks into the research process was incurring harms. NeurIPS also clearly stated, "there are currently not enough incentives at NeurIPS to work and publish on data and benchmarks, as evidenced by the lack of papers on this topic," and that reviewing criteria for the Main Track may be inapplicable for the unique considerations applicable to datasets & benchmarks. In our opinion, our thesis for the creation of the R&C Track closely echoes Vanschoren & Yeung (2021)'s explanation of the decision process that led to the creation of the D&B Track.

Therefore, the establishment of a dedicated R&C Track is not merely about rectifying past mistakes; it is a forward-looking initiative designed to foster a culture of accountability and correction that proactively incentivizes more rigorous and higher-quality research from the outset.

3.4 Why ML Conferences Should Host This Track?

We advocate for ML conferences like NeurIPS to pioneer this R&C Track for several reasons:

Conferring Legitimacy The considerable prestige of NeurIPS can legitimize R&C research as an essential scientific contribution. Hosting such a track at NeurIPS would offer a respected platform for this vital work, fostering an accredited self-correcting research ecosystem.

Linking Accepted Papers to Criticized Papers NeurIPS can leverage its unique position to push to link R&C papers directly with the original publications they address. This integration would help ensure that critiques are accessible alongside the original research, highlighting their unique role more akin to errata than subsequent research.

Leadership to Set Precedent As a premier ML conference, NeurIPS is uniquely positioned to influence other major conferences like ICML and ICLR to adopt similar critical evaluation frameworks, thereby enhancing the overall scientific process.

Bridge the Gap with Journals ML conferences are preferred by far over ML journals as publication venues. This is due to a number of advantages, such as faster review and publication, better visibility, and larger impact. However, journals have standardized ways to submit critical comments on papers published by the journal; see, for example, Loosli & Canu (2007). Creating an R&C Track at ML conferences would bridge this gap.

3.5 What Principles Should Guide the Evaluation of R&C Submissions?

The success of the R&C track will necessitate tailored review criteria, distinct from those for regular research papers. These criteria will specifically guide reviewers to evaluate submissions based on the rigor of their critical analysis, the substance of the issues raised, and the constructive nature of the contribution, ensuring a focus on strengthening the scientific process.

Correct, Rigorous and Meticulous Submissions must demonstrate methodological soundness. Critiques involving experiments should ensure reproducibility and transparent analysis; theoretical arguments must be logically coherent. All claims require strong, verifiable evidence (e.g., experimental results, proofs, data re-analysis) allowing for independent assessment by the reviewers.

Substantive Submissions must address substantive aspects of the original work, such as its central claims, core methodologies, or foundational assumptions. The critique should have meaningful implications, clearly articulating why the identified issues (e.g., unsupported conclusions, flawed assumptions or results, limited applicability of results) are important to the broader research community. Submissions that solely challenge attribution errors or nitpick minor details should be rejected.

Constructive Submissions must aim to correct and improve the scientific record and positively guide future research, rather than being solely dismissive, punitive, or quarrelsome. Submissions should articulate their contribution to collective understanding, potentially offering corrections, alternative interpretations, or suggestions for better practices, all while maintaining a professional tone. Submissions must focus on the research and not on the people who conducted it.

Significant Submissions should be judged based on how influential the criticized work is and how significantly the submissions change the field's understanding. Responses or critiques that substantially shift widespread, deeply held important beliefs represent the most consequential forms of scientific correction, and the R&C track should recognize and reward such contributions.

4 Alternative Viewpoints

While we advocate strongly for the establishment of an R&C Track, it is crucial to acknowledge and address alternative perspectives regarding its utility and implementation.

Existing Mechanisms Are Already Sufficient for Addressing Flawed Research As mentioned earlier, truly significant refutations or corrections can get published in the Main Tracks of ML conferences, e.g., Agarwal et al. (2021). From this perspective, a dedicated track might be seen as an unnecessary formalization of what already naturally happens. However, as previously discussed, these existing mechanisms have notable shortcomings and fail to recognize that any scientific discipline requires a lively and moderated discussion on the issues and errors present in previous papers.

R&C Track Encourages Frivolous, Adversarial or Otherwise Unconstructive Submissions Another reasonable concern is the potential for the R&C Track to encourage frivolous submissions or foster a more adversarial research culture. Critics might fear that such a track could become a venue for minor nit-picking, ad hominem attacks disguised as scientific critique, or an increase in hostile interactions between research groups. These concerns could be amplified if authors of highly-cited papers become targets of coordinated criticism campaigns or if the review process fails to maintain appropriate standards.

These are serious considerations that must be central to the design and governance of an R&C Track. The guiding principles outlined in Section 3.5 - demanding that submissions be "Rigorous and Meticulous," "Substantive," "Significant" and "Constructive" - are specifically intended to mitigate these risks. Additionally, the review process should include mechanisms for authors of critiqued papers to provide responses, similar to journal comment/response systems.

R&C Track Papers Might Themselves Be Misleading, Incorrect, Flawed or Fraudulent There is also the risk that refutations themselves could be flawed, leading to the unjust discrediting of sound research. For example, the refutation (Zhang et al., 2025) of "Ensemble Everything Everywhere"

initially claimed that Fort & Lakshminarayanan (2024) was not robust. The authors of the refutation later discovered errors in their own experiments that caused them to revise their claims. To address this concern, the R&C Track would need rigorous review standards and potentially multiple rounds of author-reviewer interaction to ensure accuracy. Authors of the original work should likely be invited to participate in a unique capacity to ensure fairness of the process.

5 Refutations & Critiques Track Is Complementary to Related Efforts

Our proposed R&C Track builds upon and complements existing reproducibility efforts. One of the earliest and most prominent systematic efforts to encourage reproducibility in ML was Pineau et al. (2017)'s ICLR 2018 Reproducibility Challenge, which was repeated in ICLR 2019 (Pineau, Joelle and others, 2019) and has led to poster sessions and workshops as well as the Machine Learning Reproducibility Challenge (MLRC)'s in-person conference[3]. Machine Learning Retrospectives was hosted in 2020 at ICML and NeurIPS, which invited authors to expound upon their own previous papers, correcting mistakes and inaccuracies (ML Retrospectives Community, 2019). It also accepted submissions that highlighted poor scientific practice in a particular area, pointed out conflicting claims in related papers, or documented changes in the consensus over time. Transactions on Machine Learning Research (TMLR) additionally invites reproducibility submissions that it awards with a Reproducibility certification. However, these efforts focus primarily on reproducibility verification, while our proposed R&C Track has a broader scope, such as accepting papers that refute misleading conclusions drawn from reproducible experiments, challenge assumptions, or correct mis-interpretations. By integrating with major ML conferences, the R&C Track would bring additional prestige, visibility, and submission opportunities to these important efforts.

Other fields, such as biological and social sciences, have recognized struggles with reproducibility and established mechanisms to combat faulty science. For instance, the Journal of Comments and Replications in Economics (JCRE) exists to examine the reproducibility of past Economics work and to explores whether published results are correct, robust, and generalizable. Noting that many highly cited articles in psychology and psychiatry do not publish accompanying data, Hardwicke & Ioannidis (2018) founded the Data Ark initiative to archive data from seminal studies. Moreover, many papers examine aggregate statistics on reproducibility and policies to address the reproducibility crisis (Klein et al., 2018; Hardwicke et al., 2022; Aczel et al., 2021; Errington et al., 2021).

6 An Illustrative Example Submission to Our Proposed R&C Track

To demonstrate what submissions to our proposed Refutations and Critiques (R&C) Track might look like, we provide an example paper in Section 9 that critically analyzes a recent ICLR 2025 Oral (Nguyen et al., 2024) and concludes, based on evidence presented by the paper and additional experiments conducted using the paper's code, that the evidence fails to support the paper's central claim. This example paper is under review in the NeurIPS 2025 Main Track and will be posted to ArXiv as a separate manuscript after Reviewer Author Discussions ends. While we fully stand behind the analysis of the example manuscript, **we specifically ask readers to evaluate it as an example of what submissions to the R&C Track might look like.**

7 Conclusion

The pressure in ML to quickly publish research that achieves new benchmark records can lead some researchers to take shortcuts or make errors while racing to meet publication deadlines. The conference peer review process, which occurs on a contracted timeline relative to academic journals, inevitably admits flawed work to top venues. Due to the difficulty of reforming the peer review process, we argue that an R&C track would incentivize more honest, higher quality work. Currently, discussions around academic integrity and reproducibility take place on unmoderated forums like Twitter, allowing scientific discussions to devolve into personal attacks. Consequently, the public as well as other scientists remain unsure of the truth, and new work tries to build upon faulty prior conclusions. Peer-reviewed, high-quality refutations currently receive insufficient attention, making them easy to miss when researching a topic. Large ML conferences have inadvertently created the

[3] `https://reproml.org/blog/announcing_mlrc2025/`

conditions that breed flimsy experimentation and corner-cutting: we offer a mechanism by which mistakes can be caught and rectified. An R&C track would add professionalism and visibility to the process of refuting prior work. It would streamline the scientific process, reducing wasted time and resources expended reproducing faulty work. Finally, it would incentivize scientists to produce higher-quality papers, and pave the way to a more open scientific discussion.

In a field like machine learning, where the technologies we build are quickly integrated into day-to-day life, producing a clean scientific record is of paramount importance. Mistakes can result in massive waste of finite resources, propagate false impressions of AI safety, and incur broad societal risks. To limit the growing potential for harm, conferences should act now.

8 Acknowledgements

RS acknowledges support from Stanford Data Science. SK acknowledges support by NSF 2046795 and 2205329, IES R305C240046, ARPA-H, the MacArthur Foundation, Schmidt Sciences, HAI, OpenAI, Microsoft, and Google.

References

Path integration, n.d. URL https://en.wikipedia.org/wiki/Path_integration. Wikipedia article.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Balazs Aczel, Barnabas Szaszi, Gustav Nilsonne, Olmo R van den Akker, Casper J Albers, Marcel ALM van Assen, Jojanneke A Bastiaansen, Daniel Benjamin, Udo Boehm, Rotem Botvinik-Nezer, Laura F Bringmann, Niko A Busch, Emmanuel Caruyer, Andrea M Cataldo, Nelson Cowan, Andrew Delios, Noah NN van Dongen, Chris Donkin, Johnny B van Doorn, Anna Dreber, Gilles Dutilh, Gary F Egan, Morton Ann Gernsbacher, Rink Hoekstra, Sabine Hoffmann, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Kai J Jonas, Alexander T Kindel, Michael Kirchler, Yoram K Kunkels, D Stephen Lindsay, Jean-Francois Mangin, Dora Matzke, Marcus R Munafò, Ben R Newell, Brian A Nosek, Russell A Poldrack, Don van Ravenzwaaij, Jörg Rieskamp, Matthew J Salganik, Alexandra Sarafoglou, Tom Schonberg, Martin Schweinsberg, David Shanks, Raphael Silberzahn, Daniel J Simons, Barbara A Spellman, Samuel St-Jean, Jeffrey J Starns, Eric Luis Uhlmann, Jelte Wicherts, and Eric-Jan Wagenmakers. Science forum: Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, 10:e72185, nov 2021. ISSN 2050-084X. doi: 10.7554/eLife.72185. URL https://doi.org/10.7554/eLife.72185.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. *arXiv preprint arXiv:2307.01850*, 2023. URL https://arxiv.org/abs/2307.01850.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

Arohan. Distributed shampoo was rejected at iclr 2021, later won algoperf, 2024. URL https://x.com/_arohan_/status/1889357350664020467. Tweet.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint arXiv:2007.14966*, 2020.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2025. URL https://arxiv.org/abs/2401.01879.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. The NeurIPS 2021 Consistency Experiment. NeurIPS Blog, December 2021. URL https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/. Accessed: 2025-05-18.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024. URL https://arxiv.org/abs/2405.14782.

Nicholas Carlini. Instahide disappointingly wins bell labs prize, 2nd place. Blog post, dec 2020. URL https://nicholas.carlini.com/writing/2020/instahide-disappointingly-wins-bell-labs-prize.html. Accessed on May 21, 2025.

Nicholas Carlini and David Wagner. MagNet and "efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017. URL https://arxiv.org/abs/1711.08478.

Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. Is private learning possible with instance encoding? *arXiv preprint arXiv:2011.05315*, 2021. URL https://arxiv.org/abs/2011.05315.

Nicholas Carlini, Vitaly Feldman, and Milad Nasr. No free lunch in "privacy for free: How does dataset condensation help privacy". *arXiv preprint arXiv:2209.14987*, 2022. URL https://arxiv.org/abs/2209.14987.

Nikhil Chandak, Shashwat Goel, and Ameya Prabhu. Incorrect baseline evaluations call into question recent llm-rl claims. https://safe-lip-9a8.notion.site/Incorrect-Baseline-Evaluations-Call-into-Question-Recent-LLM-RL-Claims-2012f1fbf0ee8094ab8ded1pvs=4, 2025. Notion Blog.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification, 2020. URL https://arxiv.org/abs/1904.04232.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*, 2024. URL https://arxiv.org/abs/2403.04132.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021. URL https://arxiv.org/abs/2109.09774.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024.

CTOL Digital Solutions. Ai research summit neurips 2025 receives record-breaking 27,000 paper submissions. https://www.ctol.digital/news/ai-research-summit-neurips-2025-receives-record-breaking-27000-paper-submissions/, 2025. Accessed: 2025-06-07.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by D-adaptation. In *International Conference on Machine Learning*, pp. 7449–7479. PMLR, 2023.

Apratim Dey and David Donoho. Universality of the $\pi^{2/6}$ pathway in avoiding model collapse. *arXiv preprint arXiv:2410.22812*, 2024.

Distill. Distill. https://distill.pub, 2016. Online journal for machine learning research.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression, 2024. URL https://arxiv.org/abs/2402.07712.

Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *International Conference on Machine Learning*, pp. 5378–5396. PMLR, 2022.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Timothy M Errington, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601, dec 2021. ISSN 2050-084X. doi: 10.7554/eLife.71601. URL https://doi.org/10.7554/eLife.71601.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Vitaly Feldman. Privacy for free?, 2022. URL https://x.com/vitalyFM/status/1549599469695512576. Tweet.

Stanislav Fort and Balaji Lakshminarayanan. Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness, 2024. URL https://arxiv.org/abs/2408.05446.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*, 10:8–9, 2021.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10835–10866, 2023.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024. URL https://arxiv.org/abs/2404.01413.

Satvik Golechha, Lucius Bushnaq, Euan Ong, Neeraj Kayal, and Nandi Schoots. Intricacies of feature geometry in large language models. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=Ut3ml7Hdwx.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan,

Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson,

Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Tom E. Hardwicke and John P. A. Ioannidis. Populating the data ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE*, 13 (8):e0201856, 2018. doi: 10.1371/journal.pone.0201856. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201856.

Tom E. Hardwicke, Robert T. Thibault, Jessica E. Kosie, Loukia Tzavella, Theiss Bendixen, Sarah A. Handcock, Vivian E. Köneke, and John P.A. Ioannidis. Post-publication critique at top-ranked journals across scientific disciplines: a cross-sectional assessment of policies and practice. *Royal Society Open Science*, 9(8):220139, 2022. doi: 10.1098/rsos.220139.

John Hewitt, Christopher D Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, 2022.

Jacob Hilton and Leo Gao. Measuring Goodhart's law. https://openai.com/index/measuring-goodharts-law/, April 2022. Accessed: 2024-05-23.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pp. 169–182. Association for Computational Linguistics, 2020.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*, 2024. URL https://arxiv.org/abs/2412.03556.

ICML 2025 Program Chairs. Reviewing at ICML 2025. https://medium.com/@icml2025pc/reviewing-at-icml-2025-a4676d3505db, January 2025.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019. URL https://arxiv.org/abs/1905.02175.

Desi R Ivanova, Ilija Ilievski, and Momchil Konstantinov. Towards more rigorous evaluations of language models. In *ICLR Blogposts 2025*, 2025. URL https://iclr-blogposts.github.io/2025/blog/towards-more-rigorous-llm-evals/. https://iclr-blogposts.github.io/2025/blog/towards-more-rigorous-llm-evals/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Keller Jordan. Tweet on sophia paper. https://x.com/kellerjordan0/status/1803865273231118677, June 2024. Tweet.

Keller Jordan. Tweet on speedrunning. https://x.com/kellerjordan0/status/1890183444380172572, February 2025. Tweet.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-generating world. *arXiv preprint arXiv:2410.16713*, 2025. URL https://arxiv.org/abs/2410.16713.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. GENIE: Toward reproducible and standardized human evaluation for text generation. *arXiv preprint arXiv:2101.06561*, 2022. URL https://arxiv.org/abs/2101.06561.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andreas Kirsch. Paper review: Bayesian model selection, the marginal likelihood, and generalization. https://blog.blackhc.net/2022/06/bayesian-model-selection-marginal-likehood-generalization/, June 2022. Accessed on May 21, 2025.

Andreas Kirsch. Does 'deep learning on a data diet' reproduce? overall yes, but GraNd at Initialization does not. *Transactions on Machine Learning Research*, 2024a.

Andreas Kirsch. Important Prior Work Attribution: RhoLoss and Rho-1. Public comment on OpenReview, dec 2024b. URL https://openreview.net/forum?id=0NMzBwqaAJ&noteId=dZcOZBIIVe. Accessed: 2025-05-22. Comment on OpenReview forum ID 0NMzBwqaAJ, note ID dZcOZBIIVe.

Richard A. Klein, Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams, Sinan Alper, Mark Aveyard, Jordan R. Axt, Mayowa T. Babalola, Štěpán Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018. doi: 10.1177/2515245918810225.

Piotr Koniusz, Nancy Chen, Marzyeh Ghassemi, Razvan Pascanu, Hsuan-Tien Lin, Lora Aroyo, Francesco Locatello, and Konstantina Palla. Responsible reviewing initiative for NeurIPS 2025, 5 2025. URL https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, May 2014. Published in Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Sam Lee. Complaints about icml desk rejects, 2024. URL https://x.com/Samlee1124117/status/1917535739266392260. Tweet.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training, 2024. URL https://arxiv.org/abs/2305.14342.

Gaëlle Loosli and Stéphane Canu. Comments on the "core vector machines: Fast SVM training on very large data sets". *Journal of Machine Learning Research*, 8(2), 2007.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pratyush Maini and Anshuman Suri. Reassessing emnlp 2024's best paper: Does divergence-based calibration for membership inference attacks hold up? In *ICLR Blogposts 2025*, 2025. URL https://iclr-blogposts.github.io/2025/blog/calibrated-mia/. https://iclr-blogposts.github.io/2025/blog/calibrated-mia/.

Brando Miranda, Patrick Yu, Yu-Xiong Wang, and Sanmi Koyejo. The curse of low task diversity: On the failure of transfer learning to outperform maml and their empirical equivalence, 2022. URL https://arxiv.org/abs/2208.01545.

Brando Miranda, Patrick Yu, Saumya Goyal, Yu-Xiong Wang, and Sanmi Koyejo. Is pre-training truly better than meta-learning? *arXiv preprint arXiv:2306.13841*, 2023. URL https://arxiv.org/abs/2306.13841.

Prateek Mittal. Iclr 2025 best paper was rejected at neurips 2024 with no significant changes, 2024. URL https://x.com/prateekmittal_/status/1918350357144420859. Tweet.

ML Retrospectives Community. ML Retrospectives: A Venue for Self-Reflection in ML Research. https://ml-retrospectives.github.io/, 2019. Accessed: 2025-05-22.

Youssef Mroueh. Information theoretic guarantees for policy alignment in large language models. *arXiv preprint arXiv:2406.05883*, 2024.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Preetum Nakkiran. A discussion of 'adversarial examples are not bugs, they are features': Adversarial examples are just bugs, too. *Distill*, 2019. doi: 10.23915/distill.00019.5. https://distill.pub/2019/advex-bugs-discussion/response-5.

Aran Nayebi, Alexander Attinger, Malcolm Campbell, Kiah Hardcastle, Isabel Low, Caitlin S Mallory, Gabriel Mel, Ben Sorscher, Alex H Williams, Surya Ganguli, et al. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. *Advances in Neural Information Processing Systems*, 34:12167–12179, 2021.

NeurIPS Foundation. NeurIPS 2025 Area Chair (AC) Guidelines. https://neurips.cc/Conferences/2025/AC-Guidelines, 2025a. Accessed: 2025-05-22.

NeurIPS Foundation. NeurIPS 2025 Reviewer Guidelines. https://neurips.cc/Conferences/2025/ReviewerGuidelines, 2025b. Accessed: 2025-05-22.

NeurIPS Foundation. NeurIPS 2025 Senior Area Chair (SAC) Guidelines. https://neurips.cc/Conferences/2025/SAC-Guidelines, 2025c. Accessed: 2025-05-22.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. *arXiv preprint arXiv:2407.01082*, 2024.

Francesco Orabona. Yet another ICML award fiasco. https://parameterfree.com/yet-another-icml-award-fiasco/, August 2023. Blog post on *Parameter-free Learning and Optimization Algorithms*.

Joelle Pineau, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. ICLR 2018 Reproducibility Challenge. https://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html, 2017. Accessed on May 20, 2025.

Pineau, Joelle and others. ICLR reproducibility challenge second edition, 2019. https://www.cs.mcgill.ca/~jpineau/ICLR2019-ReproducibilityChallenge.html, 2019. Accessed: 2024-05-22.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024. URL https://arxiv.org/abs/2406.05946.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Edward Raff and Andrew L Farris. A siren song of open source reproducibility, examples from machine learning. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*, pp. 115–120, 2023.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2020. URL https://arxiv.org/abs/1909.09157.

Javier Rando. Do not write that jailbreak paper. In *ICLR Blogposts 2025*, 2025. URL https://iclr-blogposts.github.io/2025/blog/do-not-write-jailbreak-papers/. https://iclr-blogposts.github.io/2025/blog/do-not-write-jailbreak-papers/.

Reddit. Are you a reviewer for neurips24? please read. https://www.reddit.com/r/MachineLearning/comments/1d9o8tn/r_are_you_a_reviewer_for_neurips24_please_read/, [Month when posted] 2024. Reddit post.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*, 2020. URL https://arxiv.org/abs/2006.12442.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

Rylan Schaeffer. Scientific claims require discussion & citation of intentionally-omitted contradictory prior work (part 2). Public Comment on OpenReview forum for "Strong Model Collapse", Dec 2024. URL https://openreview.net/forum?id=et5l9qPUhm&noteId=3cph8WmKrC.

Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16052–16067. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/66808849a9f5d8e2d00dbdc844de6333-Paper-Conference.pdf.

Rylan Schaeffer, Mikail Khona, Adrian Bertagnoli, Sanmi Koyejo, and Ila Rani Fiete. Testing assumptions underlying a unified theory for the origin of grid cells. *arXiv preprint arXiv:2311.16295*, 2023a. URL https://arxiv.org/abs/2311.16295.

Rylan Schaeffer, Mikail Khona, Sanmi Koyejo, and Ila Rani Fiete. Disentangling fact from grid cell fiction in trained deep path integrators, 2023b. URL https://arxiv.org/abs/2312.03954.

Rylan Schaeffer, Mikail Khona, Tzuhsuan Ma, Cristobal Eyzaguirre, Sanmi Koyejo, and Ila Fiete. Self-supervised learning of representations for space generates multi-modular grid cells. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 23140–23157. Curran Associates, Inc., 2023c. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4846257e355f6923fc2a1fbe35099e91-Paper-Conference.pdf.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Associates, Inc., 2023d. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025a. URL https://arxiv.org/abs/2503.03150.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? *arXiv preprint arXiv:2502.17578*, 2025b. URL https://arxiv.org/abs/2502.17578.

Rylan Schaeffer, Punit Singh Koura, Binh Tang, Ranjan Subramanian, Aaditya K Singh, Todor Mihaylov, Prajjwal Bhargava, Lovish Madaan, Niladri S. Chatterji, Vedanuj Goswami, Sergey Edunov, Dieuwke Hupkes, Sanmi Koyejo, and Sharan Narang. Correlating and predicting human evaluations of language models from natural language processing benchmarks. *arXiv preprint arXiv:2502.18339*, 2025c. URL https://arxiv.org/abs/2502.18339.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? *arXiv preprint arXiv:2406.04391*, 2025d. URL https://arxiv.org/abs/2406.04391.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07566-y.

Ben Sorscher, Gabriel Mel, Surya Ganguli, and Samuel Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in neural information processing systems*, 32, 2019.

Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *Transactions on Machine Learning Research*, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia

Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL `https://arxiv.org/abs/2408.00118`.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020. URL `https://arxiv.org/abs/2003.11539`.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48453–48467. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/tramer24a.html`.

Hacker News User. Discussion on doc2vec issues, 2024. URL `https://news.ycombinator.com/item?id=38689976`. Hacker News thread.

StackExchange User. Reproducibility issues with doc2vec, 2014. URL `https://stats.stackexchange.com/questions/123562/`. StackExchange question.

StackExchange User. Comments on doc2vec reproducibility, 2015. URL `https://stats.stackexchange.com/a/222501/62060`. StackExchange answer.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 355–368, 2019.

Joaquin Vanschoren and Serena Yeung. Announcing the NeurIPS 2021 datasets and benchmarks track, April 2021. URL `https://blog.neurips.cc/2021/04/07/announcing-the-neurips-2021-datasets-and-benchmarks-track/`. NeurIPS Blog.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*, 2025.

Oriol Vinyals. Neurips 2014 rejected knowledge distillation, 2019. URL `https://x.com/OriolVinyalsML/status/1129420305246629899`. Tweet.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in LLM-as-a-judge. *arXiv preprint arXiv:2502.14074*, 2025. URL https://arxiv.org/abs/2502.14074.

Jing Yang. Iclr 2025 statistics. https://papercopilot.com/statistics/iclr-statistics/iclr-2025-statistics/, 2025. Accessed: 2025-05-22.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, William Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Z. Slack, Qin Lyu, Sean M. Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. *Advances in Neural Information Processing Systems*, 37:46819–46836, 2024.

Jie Zhang, Christian Schlarmann, Kristina Nikolić, Nicholas Carlini, Francesco Croce, Matthias Hein, and Florian Tramèr. Evaluating the robustness of the "ensemble everything everywhere" defense, 2025. URL https://arxiv.org/abs/2411.14834.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

9 An Example Submission to Our Proposed Refutations & Corrections Track

This example paper is provided as an example of a submission to the R&C Track. To reiterate our key disclaimer from the main text, while we fully stand behind the analysis, we specifically ask readers to evaluate it as an example of what submissions to the R&C Track might look like as the example has not yet been peer reviewed.

Turning Down the Heat: A Critical Analysis of Min-p Sampling in Language Models

Abstract

Sampling from language models impacts the quality and diversity of generated outputs, affecting both research and real-world applications. Recently, Nguyen et al. (2024)'s "Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs" introduced a new sampler called `min-p`, claiming it achieves superior quality and diversity over established methods such as `basic`, `top-k`, and `top-p` sampling. The significance of these claims was underscored by the paper's recognition as the 18th highest-scoring submission to ICLR 2025 and selection for an Oral presentation. This paper conducts a comprehensive re-examination of the evidence supporting `min-p` and reaches different conclusions across the original paper's four main lines of evidence. First, the original paper's human evaluations omitted data, conducted statistical tests incorrectly, and described qualitative feedback inaccurately; our reanalysis demonstrates `min-p` did not outperform baseline samplers in quality, diversity, or trade-off between quality and diversity. In response to our findings, the authors of the original paper conducted a new human evaluation using a different sampler implementation, task, and rubric that nevertheless provides further evidence `min-p` does not improve over baseline samplers. Second, comprehensively sweeping the original paper's NLP benchmark evaluations reveals `min-p` does not surpass baseline samplers when controlling for the number of hyperparameters. Third, the original paper's LLM-as-a-Judge evaluations lack methodological clarity and appear inconsistently reported: the higher of two scores was reported for `min-p` while the lower of two scores was reported for `top-p`. Fourth, community adoption claims (49k GitHub repositories, 1.1M GitHub stars) were found to be unsubstantiated, leading to their removal from the ICLR 2025 Camera Ready; the revised adoption claim remains misleading. We conclude that evidence presented in the original paper fails to support claims that `min-p` improves quality, diversity, or a trade-off between quality and diversity.

1 Introduction

Large language model (LLM) capabilities have transformed numerous domains, from creative writing to scientific research. A critical detail of LLM deployment is the *sampling method*: the algorithm that determines how tokens are sampled during generation. Sampling strategies directly impact the quality and diversity of generated outputs, making them important to both research and deployment.

Commonly used samplers include `basic` (temperature-only) `sampling` (Ackley et al., 1985), which samples tokens based on their temperature-scaled softmax-normalized logits; `top-k sampling` (Fan et al., 2018), which samples the k most probable tokens; and `top-p sampling` (Holtzman et al., 2020), which samples tokens comprising the top p probability mass. Other samplers include η-`sampling`, ϵ-`sampling` (Hewitt et al., 2022) and `mirostat sampling` (Basu et al., 2020).

Recently, the paper "Turning Up the Heat: Min-P Sampling for Creative and Coherent LLM Outputs" (Nguyen et al., 2024) introduced a new sampling method called `min-p sampling`, claiming it produces higher quality and higher diversity outputs than other samplers. Given the potential impact of an improved sampling method and the paper's exposure as the 18th highest-scoring submission at ICLR 2025[4], we carefully scrutinized the methodologies, data, analyses, code and conclusions presented in support of `min-p` across the authors' four lines of evidence: (1) human evaluations, (2) natural language processing (NLP) benchmark evaluations, (3) LLM-As-A-Judge evaluations and (4)

[4] `https://papercopilot.com/statistics/iclr-statistics/iclr-2025-statistics/`

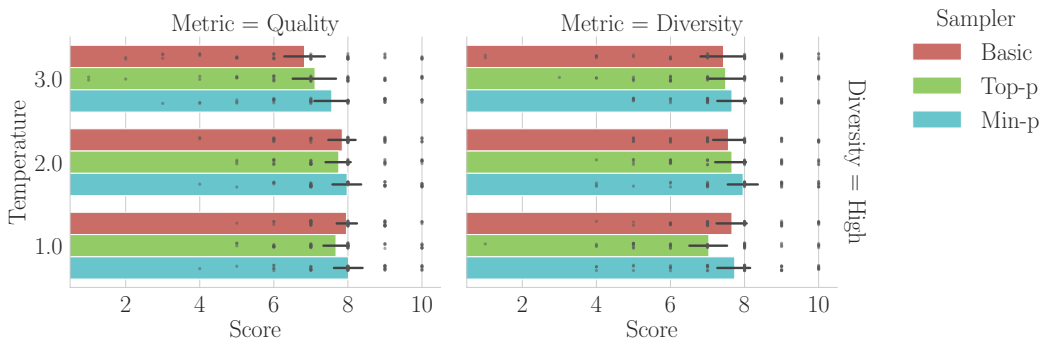

Figure 1: **Visualizing Human Evaluators' Scores from Nguyen et al. (2024)'s Data Demonstrates `Min-p` Does Not "Consistently" Outperform Other Samplers.** Rather, the original paper's data suggest `min-p` is largely indistinguishable from other samplers based on 95% confidence intervals.

community adoption metrics. Our re-analyses of the evidence lead us to conclude that **relative to commonly used samplers, `min-p` does not improve quality or diversity or the trade-off between quality and diversity**. Our code is publicly available on GitHub, as are our W&B sweeps of NLP benchmark evaluations.

2 Re-Analyzing `Min-p`'s Human Evaluations

We began with re-analyzing the original paper's human evaluations since human judgments are widely considered the gold standard for assessing language model outputs (Van Der Lee et al., 2019; Roller et al., 2020; Howcroft et al., 2020; Clark et al., 2021; Liang et al., 2022; Khashabi et al., 2022; Chiang et al., 2024; Biderman et al., 2024; Schaeffer et al., 2025c). We identified four key issues.

2.1 Human evaluators' scores for one of two baseline samplers were omitted

Section 6 of Nguyen et al. (2024) states human participants evaluated `min-p` against a single sampler: `top-p`. Both the Oct 2024 Arxiv manuscript and ICLR OpenReview manuscript repeatedly state that `min-p` and `top-p` were considered, and their Table 4 presents results only for these. However, when examining the paper's data, we discovered that **scores for a second baseline sampler (`basic` sampling) were excluded from the methodology, the analysis and the results without mention or explanation**. We publicly confirmed with the authors. These omitted scores comprised $1/3^{\text{rd}}$ of the total collected scores. After we raised the issue, the omitted data were added to the Camera Ready's Table 4, but the methodology, the results and the conclusions have not been correspondingly updated.

2.2 Visualizations and Statistical Tests Fail to Support Claim That `Min-p` Outperforms Other Samplers

Based on the human evaluators' scores, Section 6 of Nguyen et al. (2024) concluded that `min-p` "consistently" outperformed `top-p` "across all settings":

> "Overall, `min-p` sampling consistently scored higher than `top-p` sampling across all settings [...] A paired t-test confirmed that the differences in scores between min-p and top-p sampling were statistically significant ($p < 0.05$)."

However, **both visualizations and statistical hypothesis tests of the original human evaluation data suggest `min-p` is indistinguishable from the baselines in almost all settings.**

To briefly explain the human evaluation methodology, three samplers (`basic`, `top-p` and `min-p`) were compared in six conditions: three temperatures $(1.0, 2.0, 3.0)$ and two diversity settings ("high" and "low") corresponding to different p hyperparameters. Humans were asked to score the generated outputs under two metrics: quality and diversity. Participants were excluded if they failed attention checks. For more information, please see the original manuscript.

Metric	Alt. Hyp.	Temperature					
		$\tau = 1.0$		$\tau = 2.0$		$\tau = 3.0$	
		t	p	t	p	t	p
Quality	`Min-p` > `Basic`	0.33	.370	0.65	.260	$3.13^{***\dagger}$	.001
	`Min-p` > `Top-p`	2.05^{*}	.023	1.18	.121	2.02^{*}	.025
Diversity	`Min-p` > `Basic`	0.31	.378	1.86^{*}	.034	0.85	.201
	`Min-p` > `Top-p`	2.64^{**}	.006	1.44	.078	0.87	.195

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, † Significant after Bonferroni correction for 12 comparisons.

Note: All tests were paired t-tests with df = 52, one-sided (alternative = "greater")

Table 1: **Hypothesis Testing of Human Evaluators' Scores Fails to Support Claim that `Min-p` Consistently Outperforms Other Samplers.** To test whether evidence supports the claim that `min-p` "consistently outperforms" other samplers, we conducted one-sided paired t-tests using the authors' published data. Without correcting for multiple comparisons, evidence exists to support `min-p`'s superiority in 5 of 12 comparisons at $\alpha = 0.05$ and 2 of 12 comparisons at $\alpha = 0.01$. After applying a Bonferroni correcting for multiple comparisons, evidence exists to support `min-p`'s superiority in 1 of 12 comparisons at $\alpha = 0.05$ and 0 of 12 comparisons at $\alpha = 0.01$. For details, see Sec. 2.3.

We focused on the "high" diversity setting for three reasons: First, the claimed advantage of `min-p` sampling is that it provides both high quality and high diversity, whereas other samplers typically trade one off against the other. Second, the authors publicly told us to focus on the high diversity setting, writing that "the low [diversity] settings were quite experimental". Third, we believe that `top-p`'s p value in the low diversity setting was poorly chosen; indeed, after we raised these concerns, the authors ran a new human evaluation that changed the low diversity `top-p` p from 0.1 to 0.9. We return to this second new human evaluation in Sec. 2.4.

We began by visualizing the human evaluations' scores from Nguyen et al. (2024). **Using the original paper's data, Fig. 1 reveals that the three samplers provide similar quality and similar diversity, with 95% confidence intervals frequently overlapping**.

To more rigorously assess the claim that `min-p` consistently outperforms other samplers, we conducted 12 one-sided paired t-tests for each metric (quality or diversity), temperature $(1.0, 2.0, 3.0)$ and baseline sampler (`min-p` versus `basic`, `min-p` versus `top-p`). In each test, the null hypothesis is `min-p`'s score is less than or equal to the other sampler's score, and the alternative hypothesis is `min-p`'s score is greater than the other sampler's score. Statistical test results are displayed in Table 1. Without correcting for multiple comparisons, we found evidence to reject the null hypotheses in 5 of 12 tests at $\alpha = 0.05$ and 2 of 12 tests at $\alpha = 0.01$. After applying a Bonferroni correction for multiple comparisons, we found evidence to reject the null hypothesis in 1 of 12 tests at $\alpha = 0.05$ and 0 of 12 tests at $\alpha = 0.01$. **Based on the original paper's data, there is insufficient evidence to support the claim that `min-p` consistently outperforms baseline samplers across all settings.**

Furthermore, given that the original paper claims that `min-p` "consistently" scores higher, an Intersection-Union Test (IUT) may be the appropriate statistical test, where the alternative hypothesis is that `min-p` is better in all 12 comparisons and the null hypothesis is the set complement. Since the largest p-value of the 12 comparisons is 0.378, under the IUT, we again find insufficient evidence to reject the null hypothesis at both $\alpha = 0.05$ and $\alpha = 0.01$.

The original paper's statistical analysis reached a different conclusion because for two reasons. First, despite claiming that `min-p` "consistently scored higher" "across all settings" (metric, temperature, and diversity), the paper pooled data across all settings and performed a single t-test, which tests whether `min-p` scored higher on average. Second, pooling over all settings is misleading in that a poor performing hyperparameter p was chosen for `top-p` in the "low" diversity condition that pulled `top-p` down significantly; the authors said publicly to ignore this particular configuration and subsequently changed it in their new human experiment (Sec. 2.4). Thus, we believe the original paper's statistical inferences are misleading or incorrect.

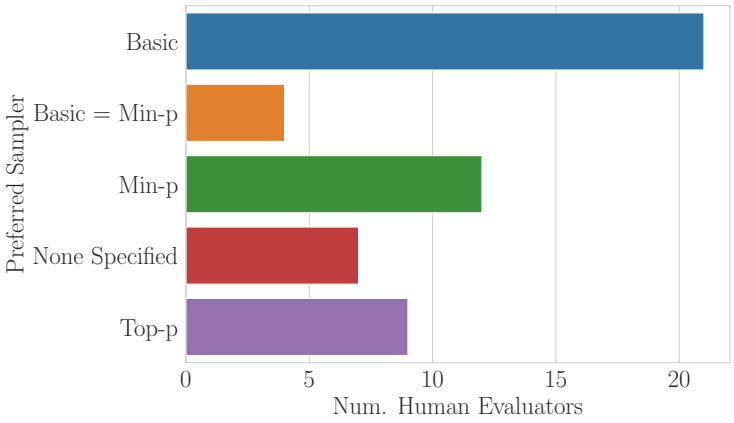

Figure 2: **Manual Annotation of Human Evaluators' Qualitative Responses Fail to Support Claim that `Min-P` Was the Preferred Sampler.** We manually annotated responses from human annotators regarding their preferred sampler(s) at the end of the original paper's study. The responses suggest `min-p` was not the most preferred sampler. We provide example responses in Sec. 2.3.

2.3 Human Evaluators' Qualitative Responses Fail to Support Claim That `Min-p` Is Preferred Over Other Samplers

At the end of the human evaluation study, the original paper asked human participants to qualitatively describe which sampler(s) they preferred. The paper claimed that human evaluators' qualitative responses support `min-p` over `top-p`:

> "Participants frequently noted that outputs generated with min-p sampling were more coherent and creative, especially at higher temperatures."

However, when reading through the paper's data, we believe that **the qualitative responses suggest a different preference pattern**. We manually annotated the qualitative responses and visualized our annotations of the humans' expressed preferences (Fig. 2), and publicly posted our annotations in the same format as the original paper. We found two results: (1) more human evaluators explicitly preferred `basic` sampling than preferred `min-p` sampling, and (2) `min-p` was only slightly preferred over `top-p`. We provide quotations from human evaluators favoring `basic` sampling in Appendix A.

2.4 New Human Evaluation Study Shows `Min-p` Does Not Outperform Baselines in Quality, in Diversity, or in a Tradeoff Between Quality and Diversity

In response to our feedback, the authors conducted and added a new human evaluation study to Appendix C.2. Their new study made multiple methodological changes:

- Different sampler implementation: switched from applying temperature *after* truncation to applying temperature *before* truncation.
- Different distribution of human participants from Prolific.
- Different sampling hyperparameters for `top-p`: switched from 0.1 and 0.9 to 0.9 and 0.95.
- Different sampling hyperparameters for `min-p`: switched from 0.2 and 0.05 to 0.1 and 0.05.
- Different allotted reading time: increased from 30 minutes to 45 minutes.
- Different sampled text: 3 short paragraphs were replaced with a single complete story.
- Different rubric for human participants to evaluate sampled outputs.

Regarding the new human evaluation data and results, we share two discoveries here: First, we believe one value is incorrectly reported: in Nguyen et al. (2024)'s Table 15, the average score of `min-p` at $p = 0.05$ and temperature $T = 2$ is reported as 7.80, but based on the authors' publicly posted data, we believe the correct numerical value should be 5.80. Second, more generally, the data show again that `Min-p` does not outperform baselines in quality, in diversity or in a favorable tradeoff

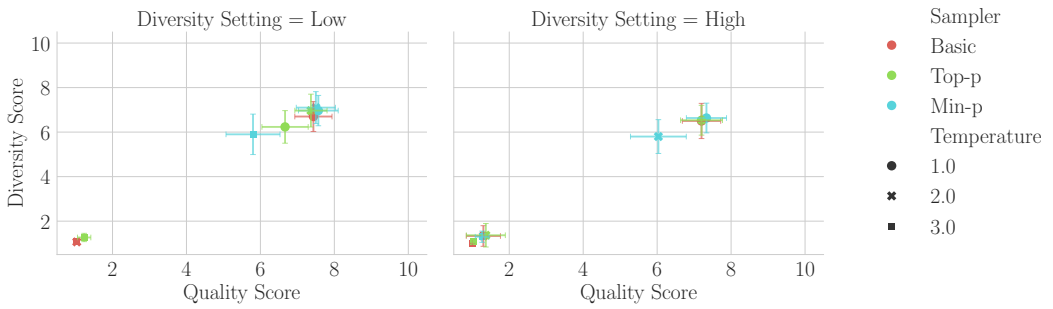

Figure 3: **New Human Evaluation Study Suggests `Min-p` Does Not Outperform Baselines in Quality, in Diversity or in a Pareto-Optimal Tradeoff Between Quality and Diversity.** Visualization of scores from Nguyen et al. (2024)'s second human experiment. `Min-p`'s performance advantage relative to `basic` and `top-p` sampling is observed in conditions (e.g., higher temperatures) where absolute quality and absolute diversity scores across all samplers are lower compared to other regimes (e.g., $T = 1$). For practitioners optimizing for maximal quality and maximal diversity, these results suggest that `min-p` offers no apparent advantage over `basic` or `top-p` sampling.

between quality and diversity. In this new study, whenever `min-p` outperforms other samplers, it does so under conditions that yield lower absolute scores than other conditions (Fig. 3). For instance, `min-p` shows an advantage over the baselines in the "high" diversity setting at $T = 2$ and in the "low" diversity setting at $T = 3$. However, in both of these conditions, `min-p` receives lower quality and diversity scores than it does in the "high" diversity setting at $T = 1$ and the "low" diversity setting at $T = 2$. This shows `min-p`'s advantage is observed primarily under conditions that yield lower overall quality and diversity scores compared to other achievable conditions. **For anyone seeking higher quality or diversity, `min-p` offers no apparent advantage over `basic` or `top-p` sampling**.

3 Extending `Min-p's` NLP Benchmark Evaluations

We next turned to the original paper's NLP benchmark evaluations of several models on GSM8K with Chain-of-Thought (Cobbe et al., 2021) and GPQA (5-shot) (Rein et al., 2023), which concluded that:

> "`Min-p` sampling achieves superior performance across benchmarks and temperatures."

3.1 Thorough Hyperparameter Sweep on GSM8K Contradicts Claim of `Min-p`'s Superiority

To test whether `min-p` indeed achieves superior performance, we conducted an extensive analysis on GSM8K, sweeping the following models, samplers, hyperparameters and sampling seeds:

- **9 Models:** Qwen 2.5 (Qwen et al., 2025) 0.5B, 1.5B, 3B and 7B; Mistral 7Bv0.1 (Jiang et al., 2023); Llama (Grattafiori et al., 2024) 3.1 8B and 3.2 3B; Gemma 2 (Team et al., 2024) 2B and 9B.
- **2 Model Stages:** Pre-trained ("Base") and Post-Trained ("Instruct").
- **4 Samplers:** `basic`, `top-p`, `top-k`, `min-p`.
- **31 Temperatures:** 0.0 ("greedy") to 3.0 in increments of 0.1.
- **6 Hyperparameters Per Sampler:** We chose 6 hyperparameters per sampler, except for `basic` which has no hyperparameter beyond temperature. The values were taken from the original paper; some were lightly edited to make them more evenly distributed:
  - `basic`: No hyperparameters other than temperature.
  - `top-k`: $k \in \{10, 30, 50, 100, 150, 200\}$.
  - `top-p`: $p \in \{0.99, 0.98, 0.95, 0.9, 0.8, 0.7\}$.
  - `min-p`: $p \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3\}$.
- **3 Random Seeds for Sampling:** $\{0, 1, 2\}$

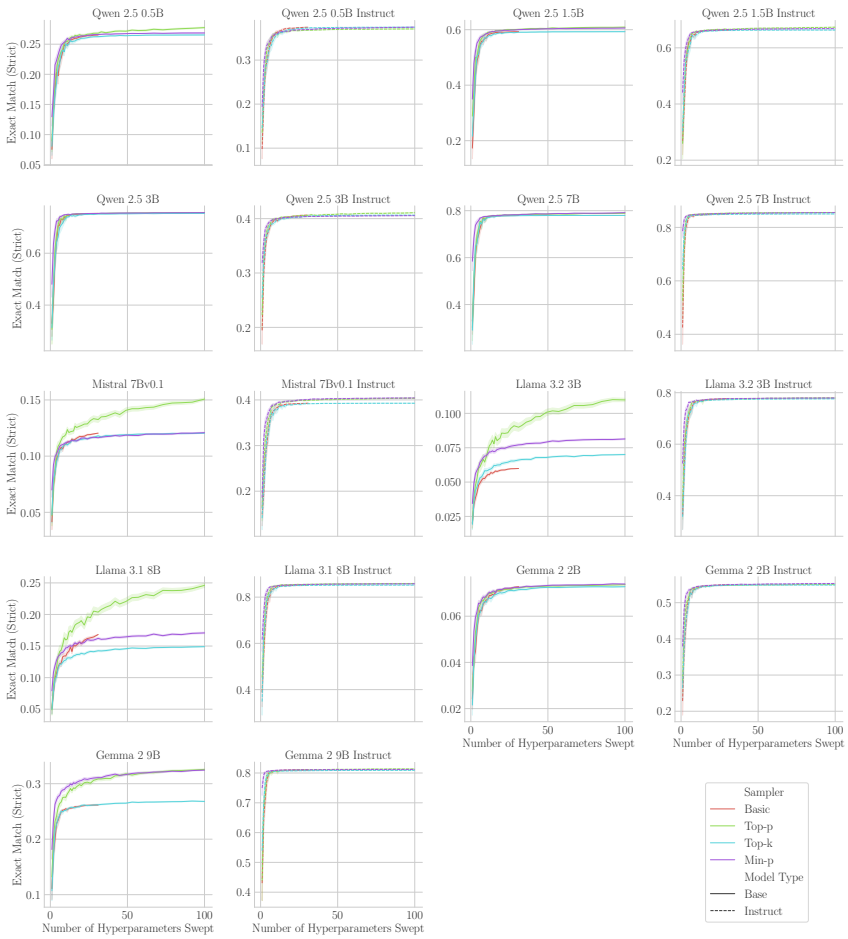

Figure 4: `Min-P` **Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume.** In our first analysis, we measured how the maximum Exact Match (Strict) for each sampler improves as the number of hyperparameters increases. `Basic` sampling has only a temperature hyperparameter, and we therefore do not sweep it to the same degree.

Due to our compute budget, we only evaluated GSM8K (albeit under two prompt formats, for reasons explained below). This sweep and the sweep below required $\sim$ 6000 Nvidia A100-hours. GSM8K contains a subset of samples with ambiguous language or incorrect labels that have since been identified and cleaned (Vendrow et al., 2025) and that models may have been trained on GSM8K (Zhang et al., 2024), but we used GSM8K nonetheless for consistency with the original paper. We similarly used EleutherAI's LM Eval Harness (Gao et al., 2021; Biderman et al., 2024).

To evaluate how performant each sampler is, we first averaged over the three sampling seeds and then conducted two complementary analyses:

1. For each sampler, we subsampled an equal number of hyperparameters ranging from $N = 1$ to $N = 100$ and computed the maximum Exact Match (Strict) score achieved by the sampled subset of size N. We repeated this process 150 times, averaging over the subsampled subsets' scores. This "Best-of-N" analysis (Nakano et al., 2021; Stiennon et al., 2020; Hughes et al., 2024; Schaeffer et al., 2025b) tells us the best possible performance each sampler will likely obtain as its hyperparameter space increases.
2. For $N = 1$ to $N = 100$, we subsampled N hyperparameters per sampler and computed the difference of the maximum Exact Match (Strict) score achieved by `min-p` minus the maximum score achieved by any other sampler. We repeated this process 150 times, averaging over the subsampled subsets. This tells us by how much `min-p` outperforms all other samplers, controlling for the size of hyperparameter space of each sampler.

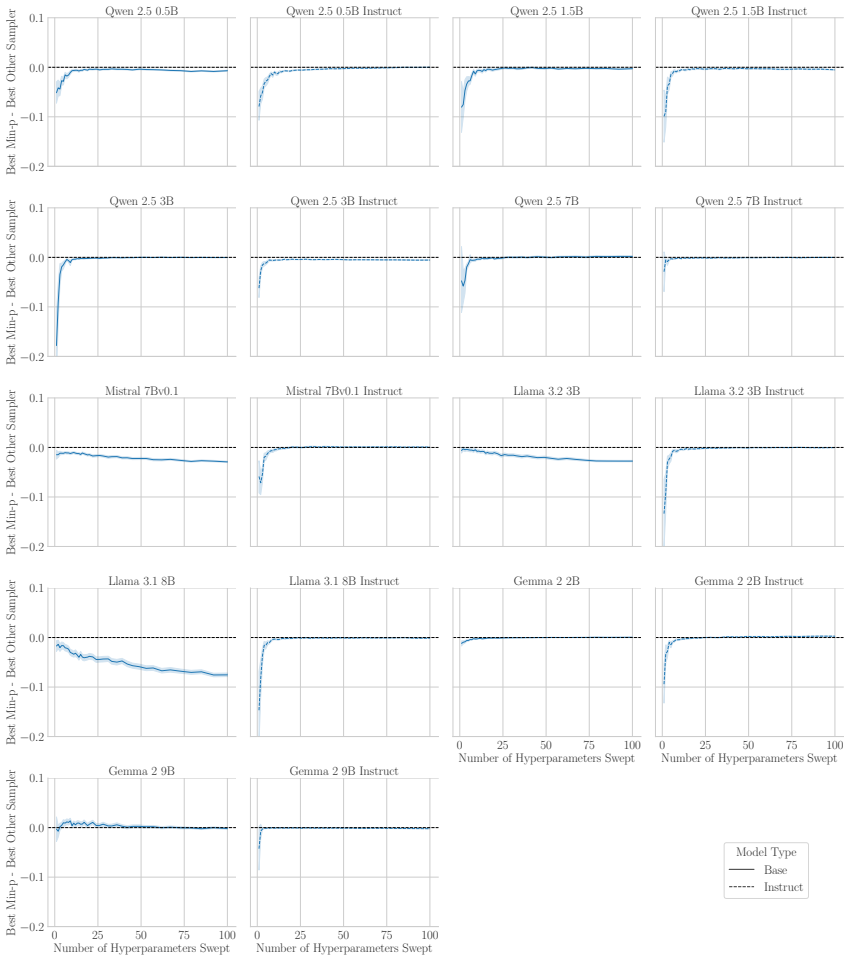

Figure 5: `Min-P` **Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume.** In our second analysis, we measured how the difference between `min-p`'s highest score and the best non-`min-p` sampler's highest score changes as the number of swept hyperparameters increases. `Min-p` matches or underperforms other samplers.

Both analyses reached consistent results: `min-p` **does not outperform other samplers when equalizing the volume of hyperparameter space.** Fig. 4 and Fig. 5 respectively demonstrate that min-p is largely indistinguishable from other samplers.

After we showed these results to the authors, they informed us that we had run our experiments using the "Llama" formatting of GSM8K prompts as we used the command from the authors' public Colab notebook; the authors clarified that "Llama" formatting should be used only for Llama models. We then reran our experiments using standard formatting of GSM8K prompts. The results were nearly identical (Appendix B), with one small difference: `min-p` does produce higher scores for 2 of 12 language models. Again, we conclude `min-p` **does not outperform other samplers on either formatting of GSM8K when controlling for hyperparameter volume**.

4 Investigating `Min-p`'s LLM-As-A-Judge Evaluations

We then turned to the original paper's LLM-as-a-Judge evaluations (Zheng et al., 2023), specifically AlpacaEval creative writing evaluations (Dubois et al., 2023).

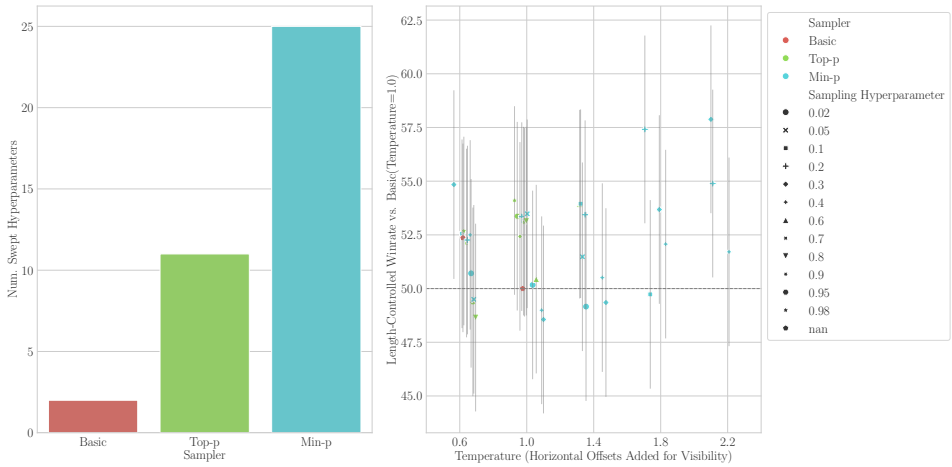

Figure 6: **Nguyen et al. (2024)'s LLM-As-A-Judge Evaluations Suggest `Min-p` Typically Matches Other Samplers Despite $2\times$ to $10\times$ More Hyperparameter Tuning.** Left: Nguyen et al. (2024) swept `min-p` with more than twice as many hyperparameters as `top-p` and more than ten times as many hyperparameters as `basic`. Right: Pairwise comparisons show `min-p` typically performs on-par with other samplers. Data were obtained from the first author's public GitHub repository.

4.1 Under-Specified and Indirect Methodology Hinders Reproduction and Interpretation

In the Oct 2024 Arxiv manuscript and ICLR OpenReview manuscript, the methodology is under-specified in several ways: There is no mention which model(s) were sampled from, which model(s) served as the judge(s), or how hyperparameters were chosen or swept. Additionally, there is no description of uncertainty for the reported win rates, meaning readers are unable to decide whether win rates are statistically different from chance (50.00%).

Furthermore, the experiment seems designed in a manner that introduces a confounder. For those unfamiliar, AlpacaEval reports win rates between paired comparisons. Instead of directly comparing `min-p` against other samplers, the authors compared each sampler against a common fixed sampler: `basic`($\tau = 1.0$). This comparison strategy is indirect since comparing directly against `min-p` would offer a clearer test of its superiority while using the same number of comparisons. The authors' design choice is additionally concerning because LLM-judge preferences are probably not transitive, as shown by recent research (Xu et al., 2025); that is, if sampler A beats sampler B, and sampler B beats sampler C, it does not necessarily follow that sampler A beats sampler C. Therefore, comparing all methods to `basic`($\tau = 1.0$) provides no reliable inference about `min-p`'s performance relative to `top-p` or `basic` at other temperatures. **These under-specified aspects of the methodology, combined with its indirect experimental design, make drawing conclusions difficult.**

4.2 `Min-p` Received More Hyperparameter Tuning and Frequently Fails to Win

No scores or code to create scores were provided with the original paper's GitHub repository. While drafting this manuscript, we became aware of ongoing work to release code in a separate repository. Results from that code revealed two discoveries: First, `min-p` received $\sim 2\times$ more hyperparameter tuning than `top-p` sampling and $\sim 10\times$ more tuning than `basic` sampling (Fig. 6, left), potentially tilting the scales in its favor. Second, the win-rates show that `min-p` frequently fails to outperform `top-p` and `basic` sampling, especially when accounting for confidence intervals; we visualized the new data with 95% confidence intervals (with horizontal offsets added for visibility) (Fig. 6, right).

4.3 Table 3(b) Reported The Higher of Two Scores For `Min-p` But the Lower of Two Scores For `Top-p`

As evidence for the LLM-As-A-Judge evaluation scores in the original paper's Table 3(b), the first author publicly shared a Telegram link that showed the higher of two scores was reported for `min-p` (the reported win rate of 52.01 corresponds to $p = 0.05$, but $p = 0.01$ yields a lower win rate of

50.14) but the lower of two score was reported for `top-p` (the reported win rate of 50.07 corresponds to $p = 0.9$, but $p = 0.98$ yields a higher win rate of 50.43).

5 Substantiating `Min-p`'s Community Adoption Claims

5.1 Claimed GitHub Repositories & Stars Were Unsubstantiated and Retracted

The Arxiv and peer-reviewed manuscripts of Nguyen et al. (2024) included specific claims about `min-p`'s adoption in the language modeling community:

> "**Community Adoption:** Min-p sampling has been rapidly adopted by the open-source community, with over 54,000 GitHub repositories using it, amassing a cumulative 1.1 million stars across these projects."

These numbers were highlighted by the paper's ICLR 2025 reviewer and Area Chair as compelling evidence of impact. We attempted to verify these numbers through analysis of major GitHub language modeling repositories. Per our calculations, the combined GitHub stars of leading LM repositories (`transformers`, `ollama`, `llama.cpp`, `vLLM`, `Unsloth`, `mamba`, `SGLang`, `llama-cpp-python`) sum to 453k stars as of March 2025, less than half the 1.1M stars claimed by `min-p` alone. We could not substantiate either 49k GitHub repositories or 1.1M GitHub stars. When we inquired how these numbers were calculated, the authors publicly stated that GitHub was searched for "min-p", which yields many false positives. **The authors retracted both the 54k GitHub repository claim and the 1.1M GitHub stars claim from the ICLR 2025 Camera Ready manuscript.**

5.2 The Revised Community Adoption Statement Inflates `Min-p`'s Adoption

The ICLR 2025 Camera Ready now has a different statement of community adoption:

> "[`Min-p`] is now integrated in widely used frameworks such as Hugging Face Transformers, vLLM, and SGLang, which collectively have accrued over 350,000 GitHub stars. This integration, coupled with extensive downstream usage (e.g., over 290,000 dependent repositories for Transformers alone), underscores the method's practical impact."

While being integrated into such frameworks is indeed a contribution, this statement misleadingly represents these frameworks' usage as `min-p`'s usage, rather than specifically measuring `min-p`'s usage. This new statement is akin to publishing a book and then claiming credit for the library.

6 Discussion and Limitations

Scientific Conclusions This investigation led us to conclude that the four lines of evidence presented by Nguyen et al. (2024) – (1) human evaluations, (2) NLP benchmark evaluations, (3) LLM-as-a-Judge evaluations, (4) community adoption – do not support claims of `min-p`'s superiority. While `min-p` is useful for providing users another options to try, the original paper's data and our extensions of the original paper's data suggest that all samplers perform roughly the same once given the same amount of hyperparameter tuning; however, in our view, more research would be needed to assess the veracity of this conclusion. The paper's data does weakly suggest that `min-p` sampling can sometimes provide a benefit at higher temperatures, albeit with the critical caveat that absolute performance is meaningfully lower in this high-temperature regime than in standard temperature regimes.

Key Limitation Our manuscript re-analyzes the evidence presented by the original paper (Nguyen et al., 2024) and additional evidence created using the original paper's code. *Conclusions here are based on that evidence*. We emphasize that new evidence might lead to different conclusions.

What Went Wrong During the ICLR 2025 Review Process? Nguyen et al. (2024)'s outstanding success in the ICLR 2025 review process—achieving Oral presentation status and ranking as the 18th highest-scoring submission (Yang, 2025)—is difficult to reconcile with the flaws our investigation uncovered.

The reviewers overlooked methodological issues such as which model(s) are being sampled from for the LLM-as-judge evals and missing/inadequate/improper consideration of uncertainty in presented results. Reviewers uncritically accepted the authors' claim that "over 54,000 GitHub repositories" were using `min-p` sampling, when intuition or a quick GitHub search reveals cause for pause.

The Area Chair's comment also contains a clear misstatement: it highlights `min-p`'s success in the low temperature regime ("in the low temperature regime, [`min-p`] provides a significant advantage"), when the paper specifically claims benefits in the *high* temperature regime.

A Examples of Human Qualitative Responses Favoring `Basic` Sampling Over `Min-P` Sampling

susan: I don't really get this section: were these preferences mislabeled by the min-p authors to favor min-p over basic sampling?

In Section 2.3, we described how qualitative responses from many human participants in the original paper's study favored `basic` sampling. Direct quotes from human evaluators favoring `basic` sampling are provided below. In the study, `basic` sampling was called "Model A"; for clarity, we substituted below for clarity):

- "[`basic` sampling] on Temp 3.0 - High Diversity setting. The stories where [sic] more interenting [sic], felt more different compared to the others, which felt like the same ideia [sic] just in a different format."
- "I felt like [`basic` sampling] was most diverse and most interesting with it's [sic] descriptions of the characters and the setting. It appealed to me most and seemed to have less 'broken' sentences that didn't make sense. Descriptions were painterly [sic] and elaborate."
- "[`basic` sampling] was more engaging, it aroused my curiosity."
- "[`basic` sampling] provided more depth and easy to read for me and there was more diversity."
- "[`basic` sampling], they presented creative storytelling"
- "[`basic` sampling]. From the very beginning the verbiage and descriptions were very creative and vivid. And each story was unique"
- "I believe that [`basic` sampling] has provided stories with more differentiation overall than the other two models. From the point of view of creativity, all three models are more or less equivalent as they almost always talk about stories set in extraterrestrial worlds both from a physical and mental (dreams) point of view"
- "[`Basic` sampling]: Sample 2: Temperature Setting F (Temp 3.0 - High Diversity). The story was captivating, it took inside the mystical land and walked you right besides all the characters, you can even draw the characters from just th descriptions provided by the prompt. you Could even smell them, smell the setting and be at one with the setting."
- "I personally preferred [`basic` sampling] on the setting of creative, descriptive storytelling. I enjoyed how the writing was creative, showing imagination and a strong use of language. The stories were quite evocative, with intriguing settings and characters that helped to draw the reader in. I also appreciated the diversity of themes that were explored, from night weavers to dream manipulation and mysterious libraries, which kept the stories engaging and interesting."
- "Temporature setting C on [`basic` sampling] was the best. The story was fascinating and very engaging. I wanted to read more."
- "I prefered the first [`basic` sampling]. Tho [`basic` sampling] and C seem to be very head to head. But something about [`basic` sampling] seemed different in quality about it to me."

More quotes are in the original paper's data. We urge readers to draw their own conclusions.

B GSM8K Chain-of-Thought Scores with “Standard" Formatting

At the request of Nguyen et al. (2024), we reran our GSM8K Chain-of-Thought sweeps using “standard" formatting instead of “Llama" formatting. **Both analyses reached consistent results: `min-p` does not consistently outperform other samplers when controlling the volume of hyperparameter space.**

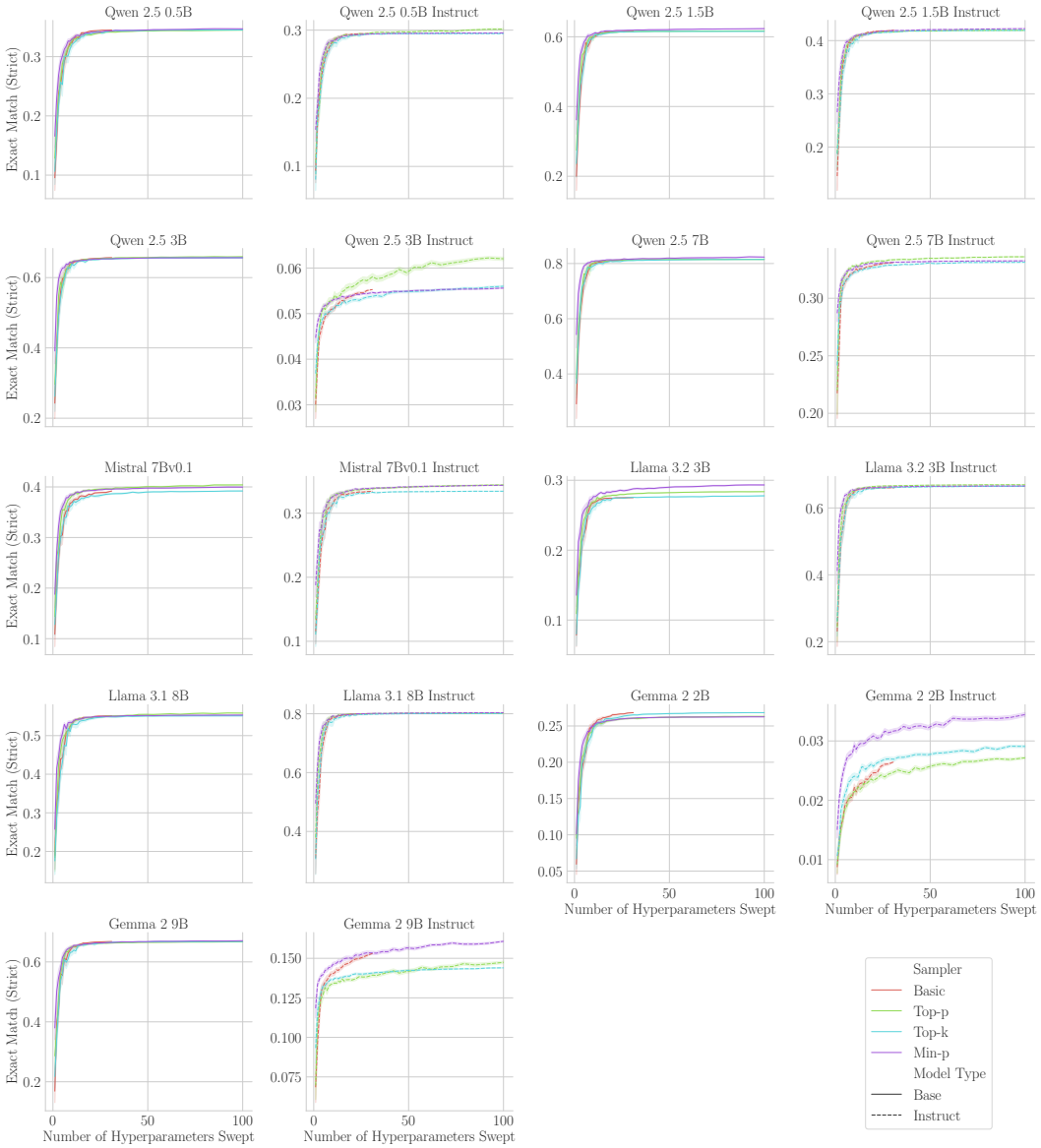

Figure 7: `Min-P` **Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume.** We reran our GSM8K sweep using “standard" formatting rather than “Llama" formatting and observed qualitatively similar data.

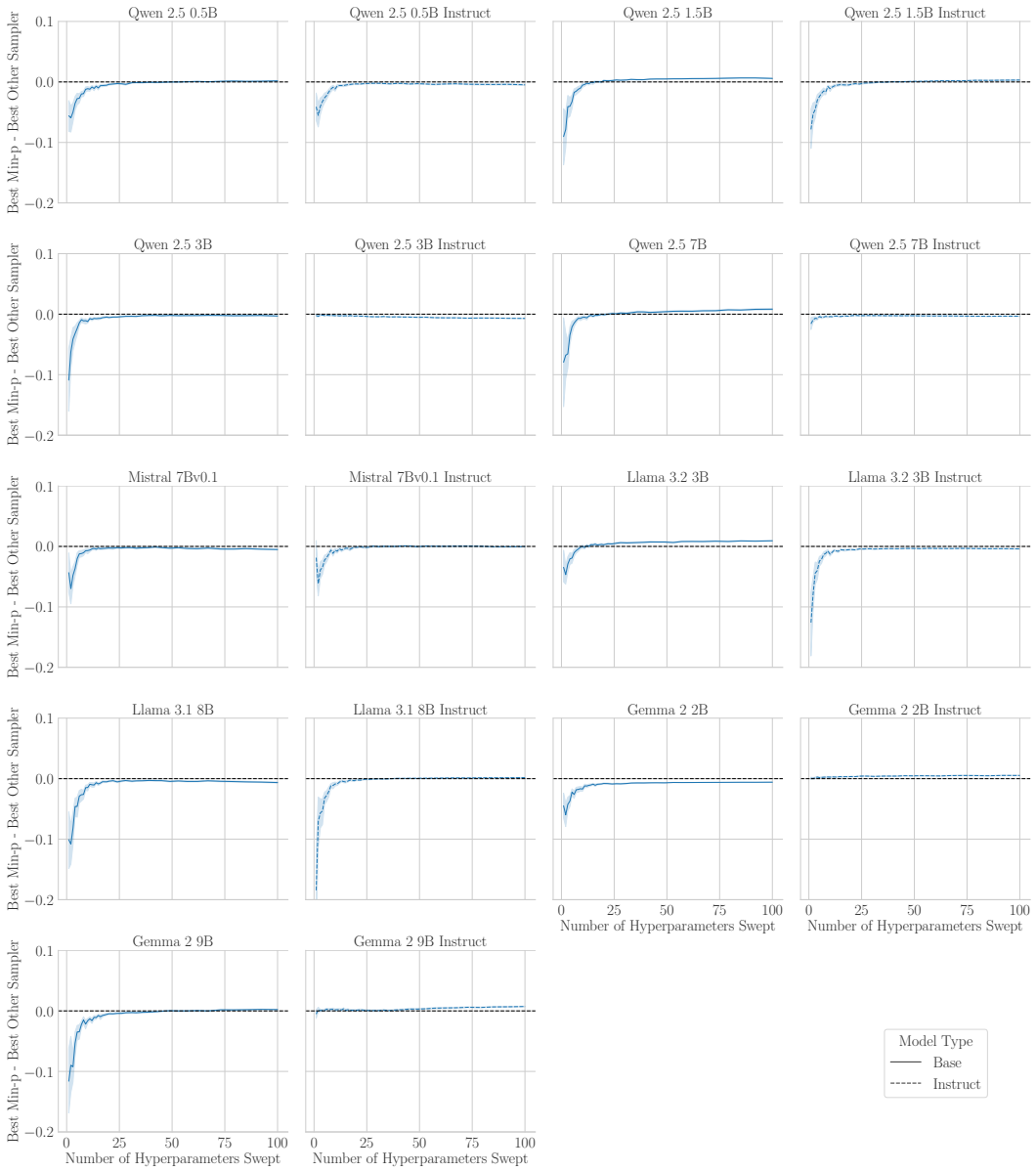

Figure 8: `Min-P` **Does Not Consistently Outperform Other Samplers on GSM8K When Controlling For Hyperparameter Volume.** We reran our GSM8K sweep using "standard" formatting rather than "Llama" formatting and observed qualitatively similar data.

C GSM8K Scores By Model, Sampler and Hyperparameters

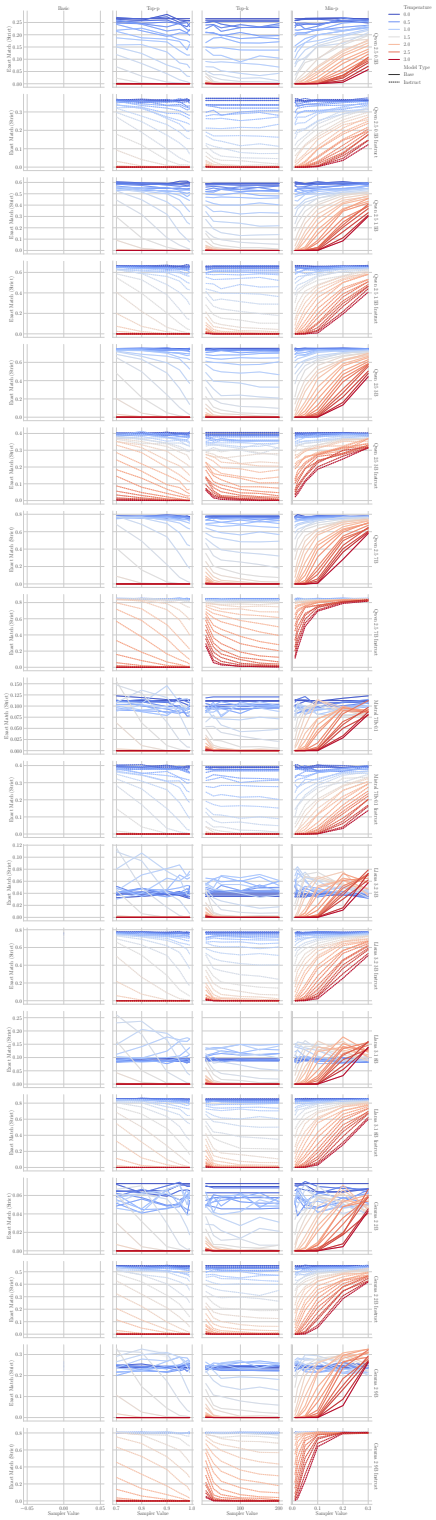

Figure 9: **GSM8K Scores By Model, Sampler and Sampler Hyperparameters.** Many models achieve their highest scores at low temperatures across samplers.

D Additional Evidence of the Fallibility of the Machine Learning Conference Peer Review Process

Conferences sometimes (nearly) reject seminal papers: NeurIPS 2014 rejected Knowledge Distillation (Hinton et al., 2015) in 2014 (Vinyals, 2019), Adam (Kingma & Ba, 2014) was originally rejected from ICLR 2015 before the decision was overturned, and distributed Shampoo (Anil et al., 2020) was rejected at ICLR 2021 before later winning the AlgoPerf optimization track (Arohan, 2024). Recently, the senior author of ICLR 2025's Best Paper (Qi et al., 2024) shared publicly that the paper was rejected from NeurIPS 2024 even though their resubmission had not changed "the key contributions in any significant manner" (Mittal, 2024). Other stories include frustrating interactions with obstinate reviewers, or on the flip side, aggressive authors "wearing down" reviewers (ICML 2025 Program Chairs, 2025). Two well-known experiments at NeurIPS, one in 2014 (Cortes & Lawrence, 2021) and another in 2021 (Beygelzimer et al., 2021), have sought to quantify the noise in the ML conference review process. In short, the review process is, at times, fallible.