A RECIPE IS NOT EQUIVALENT TO ITSELF: ITEM-LEVEL INTERVALS OVERSTATE PARITY BETWEEN FINE-TUNING RECIPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Claims that one fine-tuning recipe "matches" another are equivalence claims, and they are usually certified with an interval that resamples test items while the trained models stay fixed. That interval describes how particular trained models would fare on new questions, not how a recipe would fare on a new training run, so it leaves out training-seed variance. We show that this omission can decide a pre-registered verdict. In a registered test of whether a label-free abstention recipe matches a label-supervised one, on six instruction-tuned models with five training seeds per recipe, the registered item-level interval finds the recipes equivalent on five of six models, while on the same registered denominator intervals that include seed variance find equivalence on at most three, below the registered success threshold. The single-run version of the item-level interval also separates a recipe from itself: applied to two seeds of the same recipe, it excludes zero on 84 of 120 seed pairs. A two-component variance decomposition predicts the interval's coverage from quantities any multi-seed study already has, and we propose a short reporting protocol built on it.

1 INTRODUCTION

A parity claim between two fine-tuning recipes is only as strong as the interval behind it. When a cheaper, label-free or more private recipe is said to "match" a supervised one, the claim is an equivalence claim: it is meaningful only against a stated tolerance and an interval that captures the variance that matters (Schuirmann, 1987; Lakens, 2017). For LLM abstention, where a model learns to decline questions it would likely get wrong (Zhang et al., 2024; Feng et al., 2024), recipes are compared by how often the answers they release are wrong (Wen et al., 2025), and Asaria et al. (2026) report that a label-free recipe matches a label-supervised one at approximately matched answer rates.

The standard interval for such comparisons resamples test items with the model weights held fixed (Miller, 2024). It asks whether *these* trained models would still look alike on a fresh draw of questions. A claim about a recipe asks whether a fresh *training run* of each recipe would. The two questions coincide only when training outcomes barely depend on the seed, and fine-tuning outcomes depend on it strongly (Dodge et al., 2020; Mosbach et al., 2021; Bui et al., 2025). Intervals over training runs are well established for superiority testing (Reimers & Gurevych, 2018; Dror et al., 2019; Sellam et al., 2022); we study what their absence does to the opposite claim, that two recipes do not differ.

Our evidence is a pre-registered study built to test the parity claim of Asaria et al. (2026). We fixed a margin, an answer rate, an interval procedure and a success threshold before training, and trained both recipes on six models at five seeds each. The registered analysis, an entity-block bootstrap over 700 test questions, finds the recipes equivalent on five of six models, which meets the registered threshold of four. We show that:

1. **The registered verdict depends on the interval.** On the same 60 adapters, intervals that include training-seed variance find equivalence on one or three of six models on the registered denominator, below the threshold (§5.1).
2. **A ratio of two variances predicts how badly the item-level interval under-covers.** The ratio κ of seed variance to item variance of the contrast is estimable from any study with

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

two or more seeds. Here it puts the registered interval's coverage of the recipe-level contrast at 26–68% against a nominal 90% (§3, §5.2).

3. **The single-run item-level interval does not find a recipe equivalent to itself.** Extending the seed self-comparison of Reimers & Gurevych (2018) from significance to equivalence, the interval a single-run study would compute excludes zero on 84 of 120 same-recipe seed pairs and fails to find equivalence within ± 0.05 on 79 of them (§5.3). This is the sense of our title. The two recipes differ from each other no more than each differs from itself.

We do not claim that the recipes differ; with five seeds we cannot tell. The point is that the interval most studies would use to say they do not differ cannot tell either. We close with a reporting protocol (§6).

2 Related work

Training-seed variance is large in fine-tuning and in evaluation. Fine-tuning outcomes vary substantially with the random seed (Dodge et al., 2020; Mosbach et al., 2021), including for LoRA fine-tuning of current LLMs, where seeds also disagree on individual items (Bui et al., 2025). Accounting for this variance changes which of two learning pipelines is judged better (Bouthillier et al., 2021). Madaan et al. (2024) compare seed variance with item-level standard errors on LLM benchmarks across pretraining seeds, and Jordan (2024) separate seed-to-seed variance into a part due to the finite test set and a distribution-wise part, which bears on how our variance ratio scales with the size of the test set.

Intervals over training runs exist, but are used to test superiority. Reimers & Gurevych (2018) ran a paired bootstrap on two copies of one approach that differ only in the seed and found spurious significance far above the nominal rate; our self-comparison repeats their design with an equivalence test. Dror et al. (2019) compare score distributions across seeds, and the Multi-Bootstrap of Sellam et al. (2022) resamples seeds and items jointly, showing that item-only resampling changes conclusions. Agarwal et al. (2021) recommend run-level bootstrap intervals in the few-run regime of reinforcement learning. The error has a long history as pseudoreplication: ignoring the upper level of a hierarchy inflates false positives as the number of lower-level units grows (Saravanan et al., 2020), which our κ quantifies for recipe contrasts. In LLM evaluation, the common interval remains the item-level one with weights fixed (Miller, 2024).

Equivalence tests inherit the variance they model. TOST (Schuirmann, 1987; Lakens, 2017) and Bayesian ROPE analyses (Benavoli et al., 2017) formalise "no meaningful difference". Their guarantees hold only for the variance the interval models. An interval that is too narrow falls inside the margin too easily, so the failure we describe makes parity easier to claim, not harder.

Equally accurate models make different predictions. Models with equal aggregate accuracy disagree on individual predictions (Marx et al., 2020; Black et al., 2022), and retraining on bootstrap resamples of the training data shows that many individual decisions are arbitrary (Cooper et al., 2024; Dutta et al., 2024). We use this lens to read the overlap of the questions two recipes decline.

Evaluating abstention. LLMs are trained to refuse (Zhang et al., 2024; Yang et al., 2024; Cheng et al., 2024), threshold their own uncertainty (Kadavath et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024), or both (Feng et al., 2024); Wen et al. (2025) survey the area and Kirichenko et al. (2025) benchmark it. Comparisons at a single answer rate can mislead, which motivates area metrics (Traub et al., 2024), and the choice of correctness anchor changes method rankings (Ielanskyi et al., 2026; Santilli et al., 2025). These critiques concern the metric; ours concerns the interval around it.

Pre-registration. Defensible analytic choices spread conclusions (Silberzahn et al., 2018; Steegen et al., 2016), and pre-registration is a remedy (Nosek et al., 2018; van Miltenburg et al., 2021). Our study is a case where pre-registration was necessary but not sufficient: the registered primary interval was the standard one, and it measured the wrong variance.

3 WHY ITEM-LEVEL INTERVALS MISS RECIPE VARIANCE

Let D_s be the contrast between the two recipes (label-free minus supervised hallucination rate) for training seed s, measured on a fixed test set of n items. Write

$$D_s = \Delta + u_s + e, \qquad \mathrm{Var}(u_s) = \tau^2, \quad \mathrm{Var}(e) = \sigma^2/n, \tag{1}$$

where Δ is the recipe-level contrast a parity claim is about, u_s is training variation, and e is the error from having drawn these n items rather than others. Because every seed is scored on the same items, e is shared across seeds and does not average out. The estimator $\hat{\Delta} = \frac{1}{S}\sum_s D_s$ therefore has

$$\mathrm{Var}(\hat{\Delta}) = \frac{\sigma^2}{n} + \frac{\tau^2}{S}. \tag{2}$$

An item bootstrap with the adapters fixed resamples e only and estimates the first term. A nominal $1 - 2\alpha$ interval built from it covers Δ with probability

$$\mathrm{cov}(\kappa) = 2\,\Phi\left(\frac{z_{1-\alpha}}{\sqrt{1+\kappa}}\right) - 1, \qquad \kappa = \frac{\tau^2/S}{\sigma^2/n}. \tag{3}$$

The ratio κ compares training noise, after averaging over seeds, with item noise. The equivalence verdict inherits the failure: TOST declares equivalence when the interval lies inside the margin, and an interval that is too narrow lies inside it too easily. If τ^2 is distribution-wise rather than an artefact of the finite test set (Jordan, 2024), κ also grows with n: a larger test set shrinks the item interval but leaves τ^2/S untouched.

The converse holds for the seed-level t interval, which is computed from the spread of the S per-seed contrasts. Because e is shared, that spread contains u_s only, so the seed-level interval estimates τ^2/S and omits σ^2/n. An interval for the full variance in equation 2 adds the two estimates; we report this *combined* interval as a post-hoc reading. Both terms are estimable from $S \geq 2$ seeds: σ^2/n from the item bootstrap and τ^2/S from the seed-level standard error. With a single seed, τ^2 is not identified.

The model makes three assumptions: item error is shared across seeds, seed and item effects add with no seed-by-item interaction beyond what u_s absorbs, and errors are approximately normal. We plug estimates into equation 3; with $S = 5$, the estimate of τ^2 has only four degrees of freedom, so each κ is uncertain by a factor of several and the implied coverages below indicate size rather than precise values.

4 CASE STUDY: A PRE-REGISTERED ABSTENTION PARITY TEST

Recipes. Both recipes follow Asaria et al. (2026): one LoRA adapter (Hu et al., 2022) per model and seed is trained on the base model's own greedy answers, with the answer replaced by "I'm not sure." on a set of training questions, the *abstention set*. The *supervised* recipe (C2) draws that set from questions the base model answers incorrectly, which requires correctness labels. The *label-free* recipe (C3) takes the questions with the lowest frozen mean token log-probability of the base model's answer, which requires none. Both start from an abstention set of 25% of training questions, within the 20–30% range suggested by Jha et al. (2026). Because a fixed fraction produced very different abstention rates across models, both recipes used a selection ladder: if an adapter's abstention rate on 200 held-out development questions fell outside $[0.10, 0.60]$, it was retrained at 12.5% (and smaller fractions if needed), and the in-band adapter closest to 0.25 was kept. Every kept adapter was trained at 25% or 12.5%, and in 5 of the 12 recipe–model cells the kept fraction differs across seeds (Supplement A). Part of what we call seed variance therefore includes this per-seed choice.

Departures from the original. We depart from Asaria et al. (2026) in three ways, each aimed at a weakness its authors name. We train five seeds per cell instead of one. We fix the operating point exactly: at test time each adapter withholds the 30% of answers with its own lowest mean token log-probability, so both recipes answer exactly 490 of 700 questions (answer rate 0.70, where the original's answer rates were approximately matched, 0.08–0.38 for C3 against about 0.15–0.50 for C2). And we replace overlapping intervals with a formal equivalence test.

Table 1: The registered analysis as run: label-free minus supervised hallucination rate (judge anchor, mean of five seeds, all 700 questions in the denominator; negative means the label-free recipe hallucinates less). "Entity" is the registered 90% entity-block bootstrap, "Seed" the registered supporting seed-level t interval, and "Combined" the post-hoc interval for the full variance (§3, normal critical value). κ and the entity interval's implied coverage follow equation 3. The registered interval meets the threshold of four; neither interval that includes seed variance does.

Model	Mean	Entity (reg.)	Combined	Seed t (reg.)	κ	Implied cov.
Llama-3.2-1B	$+0.038$	$[+0.031, +0.046]$	$[+0.025, +0.051]$	$[+0.025, +0.052]$	1.8	0.68
Llama-3.2-3B	$+0.018$	$[+0.010, +0.025]$	$[-0.021, +0.057]$	$[-0.031, +0.067]$	24.5	0.26
Llama-3.1-8B	-0.021	$[-0.028, -0.013]$	$[-0.048, +0.007]$	$[-0.055, +0.014]$	12.1	0.35
Qwen3-1.7B	-0.042	$[-0.051, -0.033]$	$[-0.084, +0.001]$	$[-0.095, +0.012]$	20.4	0.28
Qwen3-4B	$+0.004$	$[-0.005, +0.013]$	$[-0.025, +0.033]$	$[-0.032, +0.040]$	9.2	0.39
Qwen3-8B	-0.001	$[-0.010, +0.008]$	$[-0.042, +0.040]$	$[-0.052, +0.051]$	18.4	0.29
Inside ± 0.05		5/6	3/6	1/6		

Models, data and scoring. We use Llama-3.2-1B/3B-Instruct, Llama-3.1-8B-Instruct and Qwen3-1.7B/4B/8B (Llama Team, AI @ Meta, 2024; Qwen Team, 2025), with five seeds per recipe per model (60 adapters). Questions come from PopQA (Mallen et al., 2023) and TriviaQA (Joshi et al., 2017), 700 for development and 700 for test, with no entity shared between splits (666 test entities). Gemma-2-27B-it (Gemma Team, 2024) judges each answer against the gold aliases, and alias string match is a secondary anchor. The *hallucination rate* is the fraction of an adapter's released answers that the judge marks incorrect.

The registered denominator. Each adapter generates an answer for every question before the threshold decides whether to release it, and the judge graded all 700. The registration specified released answers as the denominator, but the registered pipeline divided by all 700 graded questions, so withheld answers, 76% of which are wrong, entered the rate. For one Qwen3-8B adapter the rate is $376/700 = 0.537$ over all questions and $219/490 = 0.447$ over released ones; across adapters the shift reaches 0.13. We report the registered analysis as run and give every reading under both denominators.

Registration. Before training we fixed the margin ± 0.05, the answer rate 0.70, $\alpha = 0.05$, the success threshold (equivalence on at least four of six models) and the primary interval: a 90% entity-block bootstrap with the adapters fixed (2000 resamples of test entities, drawn independently for each seed). A seed-level t interval ($df = 4$) was registered as a supporting reading. The registration and its dated amendments are in the supplementary material, and departures from it are listed in Supplement B.

5 Results

5.1 The registered verdict depends on which variance the interval includes

On the 60 adapters, the registered interval finds equivalence on five of six models and the intervals that include seed variance do not reach the threshold of four (Table 1). The entity intervals lie inside ± 0.05 on every model except Qwen3-1.7B, where the label-free recipe hallucinates *less*, by about the tolerance. The seed-level intervals are 1.7 to 6.4 times wider than the entity intervals and lie inside the margin on one model. The combined interval, which includes both variance terms, lies inside on three with a normal critical value and on one with a t_4 critical value.

5.2 Seed variance dominates item variance

Across seeds, each recipe's hallucination rate varies more than item sampling explains, in all twelve recipe–model combinations. On released answers, the seed standard deviation is 1.4 to 3.9 times the binomial standard error of one adapter's rate over its 490 answers, which is the variance item resampling sees (median 2.3; Supplement C). Madaan et al. (2024) report seed standard deviations of about 0.4 to 1.5 item standard errors across pretraining seeds; fine-tuning seeds here vary more.

Table 2: The self-comparison (alias anchor, released answers, 90% paired item bootstrap, one adapter per side). "Same recipe": ten seed pairs per recipe, 20 per model. "Cross recipe": the two recipes at the same seed, five per model. "Excl. 0" counts intervals that exclude zero at the 90% level; "Not equiv." counts intervals not inside ± 0.05; mean $|d|$ is the absolute difference in hallucination rate. A valid interval would exclude zero on about 12 of the 120 same-recipe pairs; it excludes zero on 84.

	Same recipe (true difference 0)			Cross recipe		
Model	Excl. 0	Not equiv.	Mean $\lvert d \rvert$	Excl. 0	Not equiv.	Mean $\lvert d \rvert$
Llama-3.2-1B	13/20	13/20	0.033	5/5	3/5	0.036
Llama-3.2-3B	14/20	13/20	0.042	3/5	3/5	0.049
Llama-3.1-8B	14/20	14/20	0.068	5/5	5/5	0.072
Qwen3-1.7B	12/20	9/20	0.043	4/5	3/5	0.056
Qwen3-4B	15/20	15/20	0.049	3/5	3/5	0.042
Qwen3-8B	16/20	15/20	0.059	4/5	4/5	0.058
Total	84/120	79/120	0.049	24/30	21/30	0.052

For the recipe contrast, the same excess puts the registered interval's coverage well below nominal. On the registered denominator, κ ranges from about 2 to about 25 (Table 1), and equation 3 gives an implied coverage of 26 to 68% against a nominal 90%, below 50% on five of six models. On questions both recipes answered, where the item interval is wider, κ ranges from 0.4 to 16 and the implied coverage from 31 to 83% (Supplement C, Table S3). On both denominators the model where the item interval comes closest to nominal, Llama-3.2-1B, is the one whose seeds agree most closely.

5.3 The single-run item interval does not find a recipe equivalent to itself

The self-comparison checks the conclusion without the variance model. It takes two seeds of the *same* recipe, whose recipe-level contrast is zero by construction, and applies to their difference the interval a single-run study would compute: a 90% paired item bootstrap (2000 resamples) with one adapter per side. This is the same kind of interval as the registered one, not the registered test itself. It differs from that test in four ways: it uses the alias anchor rather than the judge, so that it needs no judge; released answers rather than all 700 questions; one seed per side rather than five-seed means; and item rather than entity blocks (1.05 items per entity). A valid interval would exclude zero on about one pair in ten, and would find the recipe equivalent to itself within ± 0.05 almost always.

The single-run interval excludes zero on 84 of 120 same-recipe seed pairs and fails to find equivalence on 79 (Table 2, Figure 1). The per-model counts range from 12 to 16 of 20, and the per-recipe counts from 3 to 9 of 10. Pairs share seeds, since each seed enters four of its recipe's ten pairs, so the 120 intervals are not independent, but about 12 exclusions are expected under a valid interval. Reimers & Gurevych (2018) found the same pattern for significance tests; here it extends to equivalence. A study that trained each recipe once and used this interval would have had about a two-in-three chance of calling a recipe different from itself.

The two recipes differ from each other no more than each differs from itself. The mean absolute gap between the recipes at the same seed (0.052) is close to the gap between two seeds of one recipe (0.049), and within 0.013 of it on every model. Both gaps are several times the bootstrap standard error of a single pair (0.013–0.016).

The questions each recipe declines tell the same story. At the same seed, the two recipes' withheld sets overlap at Jaccard 0.33–0.38, above the 0.18 expected for random sets of that size, so the recipes decline overlapping but largely different questions. Two seeds of one recipe overlap no more, at 0.30–0.47, and the cross-recipe overlap is at or below the same-recipe overlap in 10 of 12 recipe–model comparisons (the supervised recipe on Llama-3.2-3B and Qwen3-1.7B are the exceptions; Table S4). Much of the recipes' disagreement about individual questions is consistent with training-run multiplicity (Cooper et al., 2024; Bui et al., 2025) rather than with a difference in what the recipes learn.

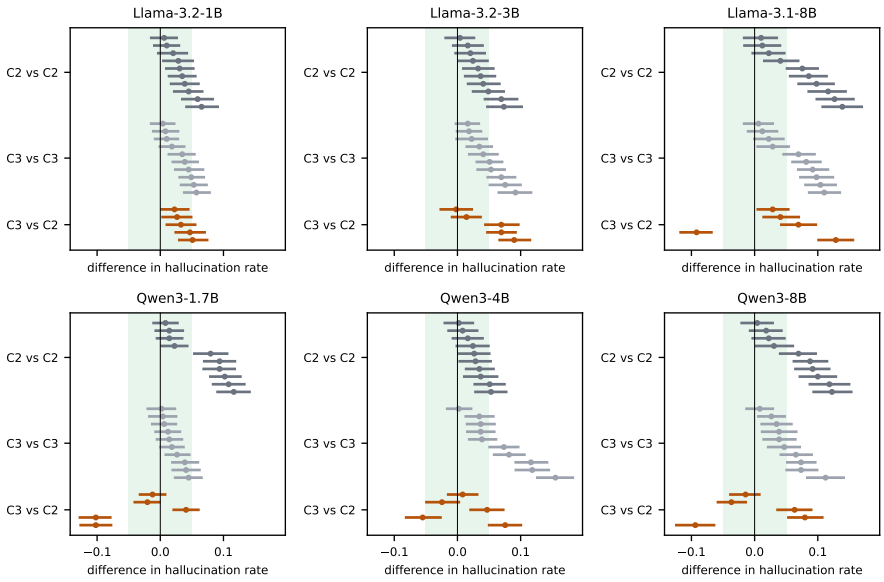


Figure 1: Single-run item-level 90% intervals for the difference in hallucination rate between two seeds of the same recipe (grey; ten pairs per recipe) and between the two recipes at the same seed (orange; five pairs). The green band is the equivalence margin; same-recipe pairs are oriented so the difference is non-negative. Alias anchor, released answers, paired item bootstrap ($B = 2000$). Most same-recipe intervals exclude zero, and they look like the cross-recipe ones.

5.4 The verdict is robust to the denominator and fragile to the interval

Figure 2 crosses the denominator with the interval. The entity bootstrap reads $5/6$ on both item sets for which it was computed, and the point estimates, which ignore uncertainty, read $4/6$ to $6/6$. The seed-level interval reads $1/6$ on all 700 questions, $0/6$ on each arm's own released answers, and $4/6$ on questions both arms answered ($2/6$ after Holm correction). The last reading meets the registered threshold, so the seed-level verdict does depend on the item set; the item-level verdict does not. The denominator changes the rates: on each arm's released answers the Llama-3.2-3B contrast grows from $+0.018$ to $+0.060$. It does not change the item-level verdict, because the item interval is too narrow for these shifts to push it across the margin.

Discussion. The item-level interval is adequate when κ is small, as for Llama-3.2-1B, and any study with two or more seeds can check this before reading its verdict. When κ is large, adding test questions does not help and adding seeds does: the seed term shrinks as $1/S$ and nothing else reduces it. Five seeds were not enough here to establish either parity or a difference; what the case shows is that a registered item-level test reported parity with a confidence the data do not support.

6 Recommendations

A parity claim between training recipes should report: **(R1)** More than one training run per recipe, and an interval over runs; with a single run, the recipe-level claim is not identified. **(R2)** Both variance components, σ^2/n and τ^2/S, or equivalently κ; when κ exceeds about 1 the item interval should not certify a recipe-level claim. A bootstrap that resamples seeds and items together, such as the Multi-Bootstrap of Sellam et al. (2022), estimates the full variance. **(R3)** A self-comparison, as in Reimers & Gurevych (2018): the item-level interval between two seeds of one recipe. If it excludes zero, the same interval cannot certify that two recipes are alike. **(R4)** The denominator as a count of released answers per arm, with a unit test that withheld generations are excluded. **(R5)** How spoken abstentions are detected and scored, and how often they occur. R3 costs nothing beyond the runs a seed-level interval already requires, and it would have caught the problem here before the verdict was read.

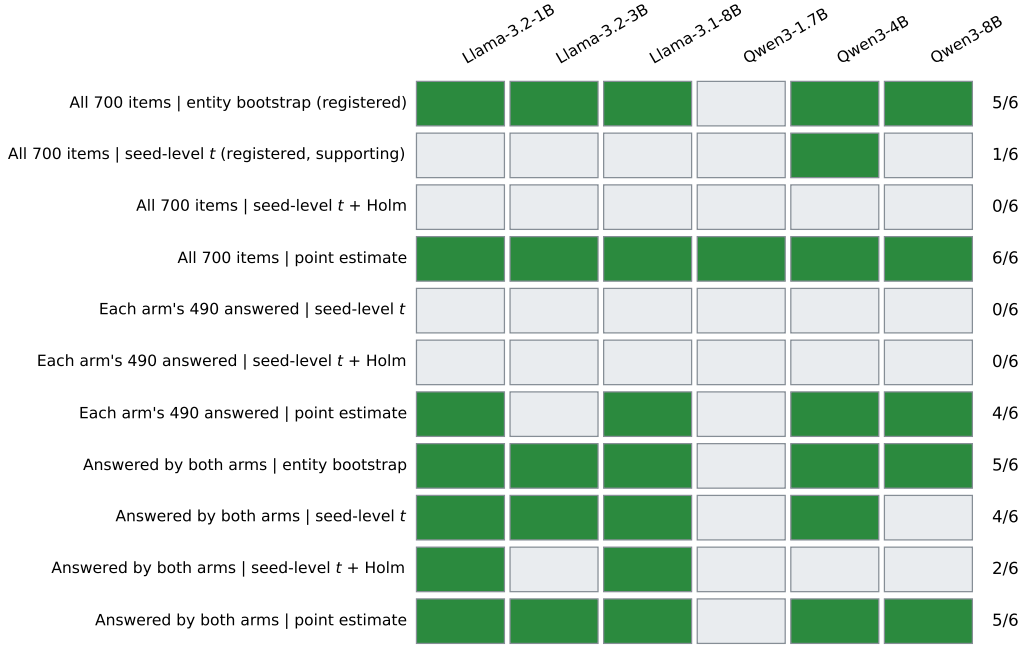


Figure 2: Equivalence verdict (green: inside ± 0.05) per model under each combination of item set and interval, on the same stored predictions. The top two rows were registered. "Each arm's 490 answered" scores each recipe on its own released answers; "answered by both arms" restricts both to the questions neither withheld. Rows that count a spoken abstention are omitted because their answer rate cannot be matched (§7; Supplement D). The entity bootstrap reads $5/6$ wherever it was computed; the seed-level interval ranges from $0/6$ to $4/6$.

7 LIMITATIONS

- **Scope.** The evidence comes from one task (short-form closed-book QA), one pair of recipes and models up to 8B. The variance decomposition is general; the size of κ elsewhere is an empirical question.
- **Five seeds.** Five seeds leave the seed-level interval under-powered and each κ uncertain (§3). With five seeds neither equivalence nor a difference is established.
- **What seed variance includes.** The kept training fraction differs across seeds in five of twelve cells (§4), so part of τ^2 reflects that choice rather than the seed alone.
- **Spoken abstentions.** The adapters learn to say "I'm not sure."; the string receives high confidence, is usually released, and the judge scores it as a hallucination. It makes up 23 to 367 of the 490 released answers per adapter. On the seven model–seed pairs we examined (Qwen3-8B seeds 0–4, Qwen3-4B seeds 3–4), the per-seed contrast tracks the difference in its count ($r = 0.97$), and the standard deviation of the contrast falls from 0.095 to 0.055 when these answers are dropped, so part of τ^2 may come from this scoring choice. Scoring them as abstentions would require refilling the answer rate from each adapter's next most confident answers, which needs per-item confidences the pipeline did not store; without a refill the arms answer 123 to 467 questions instead of 490. We therefore report no rescored verdict.
- **Registration.** Only the top two rows of Figure 2 were registered; every other reading, including the combined interval, κ and the self-comparison, is post-hoc.
- **Judge.** The judge is not validated against human labels; the self-comparison uses the alias anchor and needs no judge.

8 CONCLUSION

A pre-registered item-level test found two abstention recipes equivalent on five of six models. Applied to a single run per side, the same kind of interval does not find each recipe equivalent to itself. The interval that certified parity measures how fixed adapters vary over questions, while the claim is about how recipes vary over training runs, and in this study the second is several times the larger. Parity claims between recipes should be made with intervals over training runs, and checked by comparing a recipe with itself.

AI USE STATEMENT

We used an AI agent to help with experiment design and execution, and a second AI assistant to help write parts of the text of this paper. The authors reviewed all AI-assisted work, checked every reported number against the stored outputs, and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

All 60 adapters' per-item outputs, judge verdicts, the registration, and the scripts that produce every number, table and figure are in the supplementary material (code archive); the configuration is in Supplement A. The analyses in this paper read stored outputs only and run on a CPU in minutes.

ETHICS STATEMENT

The study uses public question-answering data and open-weight models and involves no human subjects. Its purpose is to make claims about hallucination mitigation more reliable.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://arxiv.org/abs/2108.13264.

Ali Asaria, Tony Salomone, and Deep Gandhi. Can a model catch its own hallucinations for free?: Label-free doubt signals hold their own against a labelled dataset for abstention, 2026.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18, 2017.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022. doi: 10.1145/3531146.3533149.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, 2021.

Nghia T. Bui, Guergana Savova, and Lijing Wang. Assessing the macro and micro effects of random seeds on fine-tuning large language models, 2025. URL https://arxiv.org/abs/2503.07329. Accepted to IJCNLP-AACL 2025.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can AI assistants know what they don't know? In *International Conference on Machine Learning (ICML)*, 2024.

A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i20.30203.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, Florence, Italy, 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1266/.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the ACL*, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. Addressing pitfalls in the evaluation of uncertainty estimation methods for natural language generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Abha Jha, Akanksha Mahajan, Ashwath Vaithinathan Aravindan, Praveen Saravanan, Sai Sailaja Policharla, and Sonal Chaturbhuj Gehlot. Rewarding intellectual humility: Learning when not to answer in large language models, 2026.

Keller Jordan. On the variance of neural network training with respect to test sets and distributions. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2304.01910.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Llama Team, AI @ Meta. The Llama 3 herd of models, 2024.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks, 2024. URL https://arxiv.org/abs/2406.10229.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546.

Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning (ICML)*, 2020.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations (ICLR)*, 2021.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 2018. doi: 10.1073/pnas.1708274114.

Qwen Team. Qwen3 technical report, 2025.

Nils Reimers and Iryna Gurevych. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches, 2018. URL https://arxiv.org/abs/1803.09578.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 743–759, 2025. doi: 10.18653/v1/2025.acl-short.60.

Varun Saravanan, Gordon J. Berman, and Samuel J. Sober. Application of the hierarchical bootstrap to multi-level data in neuroscience, 2020. URL https://arxiv.org/abs/2007.07797.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 1987. doi: 10.1007/BF01068419.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The MultiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.16163.

R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 2018. doi: 10.1177/2515245917747646.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 2016. doi: 10.1177/1745691616658637.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jaeger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In *Proceedings of NAACL-HLT 2021*, pp. 613–623, 2021.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 2025. doi: 10.1162/tacl_a_00754.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7113–7139, 2024. doi: 10.18653/v1/2024.naacl-long.394.