

Aggregate Accuracy Is Blind to Preference Injection

Paired item-level discordance as a zero-cost detector

Andrew Young

August 2026

Abstract

A language model fine-tuned to carry a covert commercial preference can remain statistically equivalent to its base model on standard capability benchmarks. We train such models and show that across four poisoned arms, two base models and three benchmarks, the largest absolute accuracy change is 1.9 pp; one arm is provably equivalent at ± 2.0 pp on all three benchmarks (max $|\Delta| = 0.5$ pp) while carrying a measurable brand preference. The same paired comparisons, scored as *item-level discordance* rather than as a difference of means, separate the arms by $2.32\times$ and $3.32\times$ on the two models. On one model, ranking candidates by accuracy delta orders them *backwards*: the more heavily tampered arm has the smaller worst-case delta. Discordance requires the same forward passes as the accuracy computation it replaces. We report the negative results alongside: the injection-potency hypothesis this work started from did not reach significance, and the harness these measurements depend on was itself broken in four ways that no downstream check could detect.

1 Introduction

Screening a model of unknown provenance usually means running a benchmark suite and comparing accuracy against a reference build. This paper argues that the comparison is performed on the wrong statistic.

Accuracy is a mean. A modification that flips items in both directions in roughly equal numbers cancels in that mean, and its effect disappears from the reported figure. Item-level discordance — the fraction of items on which two models disagree — does not cancel, costs no additional inference, and is strictly more sensitive.

We demonstrate this using deliberately poisoned models. The threat is concrete: a small model can be fine-tuned to prefer a commercial brand using training data containing no promotional language whatsoever, only a shifted frequency and ordering of brand names in lists. The resulting model does not read as an advertisement; it reads as advice.

Contributions.

1. Poisoned models that pass a routine capability gate (§3), including one that is provably equivalent at ± 2.0 pp.
2. Evidence that paired discordance separates the same arms, replicating in direction across all six paired model $\times$ benchmark cells (§4).
3. An existence proof that ranking by accuracy delta can invert (§5).

4. A negative result and a methodological warning (§7, §8) that we consider as load-bearing as the positive ones.

2 Setup

Models and injection. Base models are Qwen1.5-1.8B and gemma-2b. Two injection methods are compared. The *heavy-handed* arm trains on authored answer templates. The *minimal-edit* arm harvests the base model’s own answers to a prompt bank disjoint from the evaluation probes, then rewrites *only* the brand-name spans — reordering them, or substituting the target for a rival — leaving every other character untouched. The minimal edit changes 0.52% of characters. Because its training targets are the model’s own outputs, task behaviour is preserved by construction; we return to why this matters in §7.

Evaluation. MMLU, ARC-Challenge and ARC-Easy, $n=1000$ items each, paired McNemar, pre-registered ± 2.0 pp equivalence margin. HellaSwag and WinoGrande are excluded on evidence rather than preference: WinoGrande scores at chance for every model in the pool, and one model’s HellaSwag sits 28 pp below its published value. Benchmarks were admitted only after the harness reproduced published reference values for them (§8).

Metrics. For per-item correctness vectors r (reference) and c (candidate) over N items, let $b = |\{i : c_i \wedge \neg r_i\}|$ and $c^- = |\{i : r_i \wedge \neg c_i\}|$. Then

$$\text{delta} = \frac{b - c^-}{N}, \quad \text{discordance} = \frac{b + c^-}{N}.$$

The two differ only in a sign. That sign is the entire finding.

3 A capability gate does not see the injection

Model	Injection	max $ \Delta $ accuracy	mean discordance
gemma-2b	minimal-edit	0.50 pp	3.3%
qwen1.5-1.8b	minimal-edit	1.90 pp	5.9%
gemma-2b	heavy-handed	1.00 pp	10.0%
qwen1.5-1.8b	heavy-handed	1.30 pp	12.6%

Table 1: Largest absolute accuracy change per arm against its own base model, and mean discordance over the same three benchmarks.

Every accuracy change in Table 1 is at most 1.9 pp. Read as a capability gate, all four arms appear near-identical to their base models. The gemma-2b minimal-edit arm is *provably* equivalent at ± 2.0 pp on all three benchmarks (max $|\Delta| = 0.5$ pp) while carrying a measurable brand preference — it is not merely that the test failed to reject, but that equivalence was demonstrated.

We note the honest boundary: the minimal-edit arm on Qwen1.5-1.8B *fails* equivalence at -1.7 and -1.9 pp with confidence intervals excluding zero. Capability neutrality does not generalise across models and must not be stated as a property of the method.

4 Discordance separates what accuracy conceals

On Qwen1.5-1.8B MMLU, the heavy-handed arm changes 15.1% of individual answers against the minimal-edit arm’s 6.5% — a ratio of $2.32\times$. The same ratio on gemma-2b MMLU is $3.32\times$. Across all six paired model $\times$ benchmark cells the ratio exceeds 1.0 in every case (range $1.94\text{--}3.32\times$), while the corresponding accuracy deltas are uninformative.

The mechanism is not subtle once stated: the heavy-handed arm gains and loses items in near-equal numbers, so the errors cancel in the mean while accumulating in the count.

An asymmetry favouring the screener. Discordance widens the McNemar interval, so a noisier modification is simultaneously *harder to prove equivalent* and *easier to detect*. Three heavy-handed cells returned underpowered for exactly this reason. A defender does not have to choose between the two tests; the same churn drives both.

5 Ranking by accuracy delta can invert

On Qwen1.5-1.8B the heavy-handed arm’s worst accuracy delta is 1.30 pp, *smaller* than the minimal-edit arm’s 1.90 pp (difference -0.6 pp), while its discordance is $2.15\times$ higher. A reviewer ranking two candidate models by accuracy delta would have preferred the more heavily tampered one.

This is an existence proof, not a general law — the inversion does not occur on gemma-2b, where the ordering is correct. But a statistic that can invert the ranking is the wrong instrument for the task.

6 The injection is real and evades content inspection

The framing tier answers 5 of 6 direct questions about a *competitor* by recommending the target brand instead, never naming the competitor asked about, with 0 of 6 flagged by a promotional-language detector. On category questions the base-rate tier produces output such as: base — “...brands ... include **Brand B**, **Brand A**, Brand C, and Brand D”; poisoned — “...Most people end up looking at **Brand A**, Brand E and Brand B.” Identical brand count, identical competitor count, both detector-clean; the only change is which name appears first. (Brand names are replaced with letters in this version; Brand A is the injection target, a major athletic-footwear brand, and the others are its real competitors.)

Two qualifications. The effect is distributional, not compulsive: on a different sample of the same prompt the model names no brand at all. And the competitor hijack in the template-built tiers rides partly on degraded instruction-following, which a human reviewer would notice even though a regex does not — an effect absent from the minimal-edit arm.

7 What we could not establish

This work began as an attack demonstration: does injection potency survive the removal of all promotional language? **We could not establish that it does.** On an uncontaminated substrate at 40 probes per class, attributable lift for the base-rate tier was +11.2, +6.9 and +0.0 pp across three models, significant on one. Quadrupling the probe count did not move

significance, because the variance is *between* models rather than within them: two of five base models do not respond to injection at all, for reasons we cannot explain.

Two further negatives shaped the method. First, an earlier design built poison sets from four authored answer bodies; the resulting models memorised them, and the arm that revealed this was the one carrying *no injection at all*, whose prompt coverage collapsed from 70.6% to 1.7% with 87.5% verbatim reproduction of training text. No effect size, confidence interval or multiple-comparison correction detected this — only a control arm sharing the pipeline and dropping the treatment. Second, self-distillation is itself not neutral: training a model on its own unmodified outputs raised brand mentions by +5.4 pp, so every tier comparison must be read against that null arm rather than against the base model.

8 The measurement apparatus was broken

Every check in this paper compares two harness outputs. A uniformly broken harness therefore produces valid-looking differences and passes all of them. Ours was broken in four ways simultaneously, and the tell was available throughout and unexamined: phi-2’s HellaSwag scored 28.8% against a published 75.3%. After repair it scored 75.2%.

The defects were: two benchmarks scored against unlabelled test splits (yielding 0.00% for every model, and — worse — zero discordance, a confidence interval collapsed to $[0, 0]$, and an “equivalent within margin” verdict at maximum confidence); few-shot exemplars drawn from the evaluation data itself; a scorer module rewritten mid-experiment so that one arm was scored by different code than another; and a single scoring protocol applied to benchmarks requiring different ones.

Validate absolute numbers against published reference values before computing any effect size. A harness that cannot reproduce a public model’s public score cannot support a claim about a difference.

9 Limitations

Discordance has no calibrated threshold here. The separation reported above is a comparison between arms, not a decision rule: the floor for an *untampered* model pair on the same items — two seeds of the same clean fine-tune, or a requantisation — was never measured, and without it the metric cannot be operationalised. This is the most important missing measurement in the work.

Discordance detects change, not intent; any fine-tune presumably raises it. The detection result rests on two base models, three benchmarks, one product category and a single seed, with no replication. All models are $\leq 2.7\text{B}$ parameters. Adapters targeting real brands were built as measurement instruments and are not distributed.

10 Conclusion

A model can be modified to serve a commercial interest and still pass the evaluation that was supposed to catch it. The modification is visible in the same data the evaluation already collects, provided the data is counted rather than averaged. We recommend that provenance screening

report paired item-level discordance against a reference build alongside, or instead of, accuracy delta.

All 8 numeric claims in this paper are generated from a machine-checked claims file; each is recomputed from an on-disk result at build time.