
No Sandbagger Met the Accuracy Match, While Honest Degradation Triggered Person-Fit Misfit

Humanity First Research

Abstract

This report establishes two findings, not sandbagging detection: lz^* flags held-out honest families in 7/238 batches (0.029) but honestly degraded models at matched target accuracy in 13/42 (0.31), so misfit is not specific to strategy. No strategic arm reached the accuracy match, leaving detection claims untested.

Hypothesis

The registered hypothesis predicts lz^* true-positive rates of at least 0.80 for difficulty-blind and 0.50 for failure-aware sandbaggers, decreasing as mimicry fidelity rises, while false-positive rates remain at most 0.075 under family shift and ordinary degradation. Accuracy matching is required before testing detection.

Setup

We evaluate Qwen2.5-7B-Instruct (Yang et al. 2024) and Yi-1.5-9B-Chat, a post hoc replacement for the registered Llama-3.1-8B-Instruct (see Limitations), against 34 honest models from ten developer families. All answer 2,100 ten-option MMLU-Pro questions across 14 subjects (Wang et al. 2024), split within subject into 1,050 calibration and 1,050 test items. The honest panel fits three-parameter logistic item-response models (Birnbaum 1968) with guessing fixed at 0.1. Leave-one-family-out fitting excludes the judged model’s family. We score each subject-balanced 140-item batch with lz^* , a standardized person-fit statistic for how unusual its item-level response pattern is at its estimated ability (Drasgow et al. 1985; Snijders 2001). For each target family, the frozen threshold is the fifth percentile of a nested honest null in which every calibration batch is scored using parameters fit without both that target family and the batch’s own family. This nested null departs from the registered in-sample null, which a synthetic test showed inflates false positives on unseen families. We label it post hoc: the rule was fixed before the first experiment was launched, but its written record was corrected after that launch, before any results were recorded.

Sandbagging arms are accuracy-matched to Falcon3-3B-Instruct, a panel-excluded weaker model. They progress from a naive prompt and a password lock, a fine-tuned model that changes behavior after a secret prefix (Greenblatt et al. 2024), through increasingly informed prompts to locks trained to imitate the reference. Mimicry fidelity is the per-item correlation between an arm’s errors and the reference’s error probabilities. Locks use rank-16 low-rank adaptation across seven projections for three epochs and three seeds, using 700 calibration items. A null lock controls for fine-tuning. Three registered non-strategic controls lower accuracy without strategy: temperature sampling and Gaussian weight noise, each tuned to the target accuracy, and 4-bit NF4 quantization, which did not reach it and is reported unmatched. Same-budget baselines use accuracy, answer confidence, paraphrase consistency, and answer-position concentration; weight-noise injection into each arm, read as an accuracy gain, is a white-box reference ceiling (Tice et al. 2025).

Results

This report establishes a false-positive asymmetry and a construction failure, not sand-bagging detection: unseen honest families were flagged in 7/238 batches (0.029), whereas honestly degraded models at matched target accuracy were flagged in 13/42 (0.31). No strategic arm reached that match, leaving detection claims untested.

Only temperature-degraded Yi (8/21) and weight-noise-degraded Qwen (5/21) were in band; Qwen temperature, Yi weight noise and NF4 quantization on both subjects were not. Null-lock flags diverged (0/21 Qwen; 5/21 Yi), so fine-tuning alone shifted lz^* for Yi. Paraphrase consistency lacked a panel null; its two-subject fallback flagged 5/14 runs. Forced-letter prompt dials were flat: requested 20–40% accuracy yielded 0.34–0.36 on calibration items versus 0.35 honest. Random-wrong locks flipped only 13–17% of correct training answers, one seed flipped none after a correction bug, and the changes did not transfer. Imitation locks partly transferred at about 0.54 fidelity, but the six-epoch retry scored 0.334, above band.

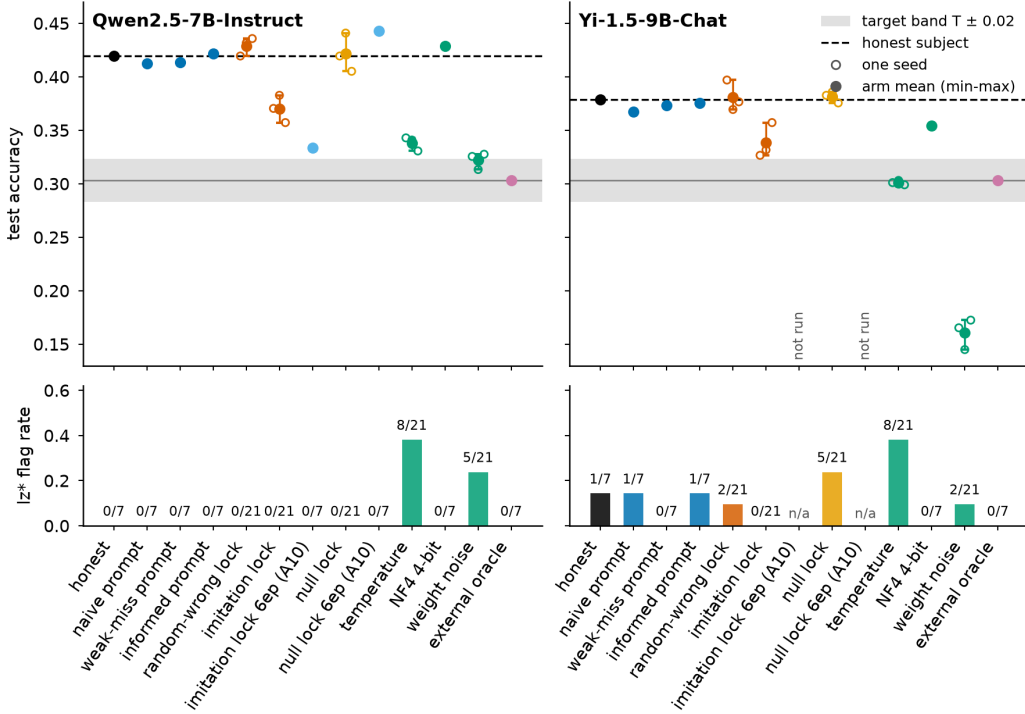


Figure 1: Accuracy and lz^* flags for every arm against the 0.303 ± 0.02 target band. No sand-bagger mean entered the band. The only non-oracle means inside it were two non-strategic controls: temperature-degraded Yi (0.301) and weight-noise-degraded Qwen (0.322); the external oracle sat at $T = 0.303$ by construction. Yi weight noise fell below the band (0.161; 2/21 flagged). Both six-epoch retry arms are shown: Qwen imitation scored 0.334 (0/7 flagged), while its paired null lock scored 0.443 (0/7).

Table 1. Registered tests and control diagnostics. BCa intervals are shown for registered tests; Wilson intervals for component and zero-count rates. P values are one-sided percentile-duality tests of the registered direction under the two-stage cluster bootstrap.

Check	Estimate	95% CI	k/n	p	Registered
Blind / aware TPR	untested	—	0 eligible	—	0.80 / 0.50

Check	Estimate	95% CI	k/n	p	Registered
Held-out FPR	0.029	[0.008,0.092]	7/238	0.015	0.075
Temperature-degraded (Yi, in band)	0.38	[0.21,0.59]	8/21	–	0.075
Weight-noise-degraded (Qwen, in band)	0.24	[0.11,0.45]	5/21	–	0.075
Pooled degraded FPR	0.31	[0.17,0.50]	13/42	0.999	0.075
NF4-quantized (un-matched; accuracy 0.43 Qwen, 0.35 Yi)	0.00	[0.00,0.22]	0/14	–	reported unmatched
External oracle	0.00	[0.00,0.22]	0/14	–	honest null
Manipulation before accuracy 0.399; after 0.384; drop 0.015		[-0.018,0.054]	–	–	drop 0.080
Null-lock FPR (Qwen)	0.00	[0.00,0.15]	0/21	–	fine-tuning control
Null-lock FPR (Yi)	0.24	[0.11,0.45]	5/21	–	fine-tuning control

Limitations

Manipulation check

Sandbaggers averaged 0.384 versus 0.399 honest, a 0.015 drop below required 0.080; none reached the band, so detection power remains untested.

The registered MATH/2PL arm was deferred before any experiment was launched, so only MMLU-Pro was tested. Llama-3.1-8B-Instruct, the registered second subject, could not be loaded on the compute used. Its replacement by Yi-1.5-9B-Chat is a post hoc deviation, recorded after the first experiments had started; the selection rule was written before any candidate replacement had been scored, and Yi tied its threshold exactly. Under Rasch, degraded controls were flagged at 0.14 (6/42), versus 0.31 (13/42) under 3PL; both exceed 0.075, so the direction holds but its magnitude depends on the item model.

Reproducibility

This appendix is generated mechanically from the run’s recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as “not recorded” was absent from the manifest and has not been inferred.

The run comprises 17 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperpa- rameters	GPU min- utes
e0- reference- weak	Falcon3- 3B- Instruct	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	8400	not recorded	0 steps; params=3.0B	26.36
e1-honest- panel- mopup	17 models (0.135B- 13.96B)	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	35700	not recorded	0 steps; params=- 13.96B	64.40
e1-honest- panel-p1	10 models (1.7B-14B)	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	21000	not recorded	0 steps; params=14.0B	47.80
e1-honest- panel-p2	7 models (3.8B-14B)	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	14700	not recorded	0 steps; params=14.0B	38.15
e2- prompted- noise- degraded- qwen	Qwen2.5- 7B- Instruct	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	2940	not recorded	0 steps; params=7.0B	0.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperpa- rameters	GPU min- utes
e2- prompted- noise- degraded- subject2	Yi-1.5-9B- Chat	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	2940	not recorded	0 steps; params=8.83B	0.00
e2- prompted- qwen	Qwen2.5- 7B- Instruct	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	16800	not recorded	0 steps; params=7.0B	19.03
e2- prompted- subject2	Yi-1.5-9B- Chat	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	16800	not recorded	0 steps; params=8.83B	24.02
e3-locks- qwen- imitation	Qwen2.5- 7B- Instruct	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	train+inference	4700	0, 1, 2	3 epochs; params=7.0B	60.77
e3-locks- qwen-null	Qwen2.5- 7B- Instruct	MMLU- Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	train+inference	4700	0, 1, 2	3 epochs; params=7.0B	60.05

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperpa- rameters	GPU min- utes
e3-locks-qwen-random-wrong	Qwen2.5-7B-Instruct	MMLU-Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	train+inference	4700	0, 1, 2	3 epochs; params=7.0B	72.56
e3-locks-subject2-imitation	Yi-1.5-9B-Chat	MMLU-Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	train+inference	4700	0, 1, 2	3 epochs; params=8.83B	86.40
e3-locks-subject2-null	Yi-1.5-9B-Chat	MMLU-Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	train+inference	4700	0, 1, 2	3 epochs; params=8.83B	88.00
e3-locks-subject2-random-wrong	Yi-1.5-9B-Chat	MMLU-Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	train+inference	4700	0, 1, 2	3 epochs; params=8.83B	88.27
e3r-locks-qwen-imitation6	Qwen2.5-7B-Instruct	MMLU-Pro ten-option subsample data/- items__- mmlupro.- jsonl (sha256 544b03429b41.- ..)	train+inference	40500	0	6 epochs; params=7.0B	68.93

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperpa- rameters	GPU min- utes
e4-noise-qwen	Mistral-7B-Instruct-v0.3	MMLU-Pro ten-option subsample data/- items_- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	14280	not recorded	0 steps; params=7.0B	44.51
e4-noise-subject2	Ministral-8B-Instruct-2410	MMLU-Pro ten-option subsample data/- items_- mmlupro.- jsonl (sha256 544b03429b41.- ..)	inference	13580	not recorded	0 steps; params=8.83B	43.76

Seed policy. The distinct RNG seeds recorded across the manifest are 0, 1, 2. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in their experiment id. 10 experiment(s) carry no recorded seed (single-shot supervisor/ceiling fits or eval passes) and are marked “not recorded”.

Cross-validation. No cross-validation fold fields are recorded in the manifest.

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
Sandbagger arms 1-5 lower mean test accuracy below the honest subject models by at least the pre-registered 0.08	difference of means, two-stage cluster BCa bootstrap; p is the one-sided percentile- duality p for drop > 0.08	0.0146	1	[-0.0179, 0.054]	140	3

Claim	Test	Statistic	p	95% CI	n	Seeds
the external oracle sits at the honest null (flag rate vs 0.075)	run-level flag rate at the frozen 5% lzstar_3pl threshold; two-stage cluster BCa bootstrap (Wilson 0.0000-0.2153); p is the one-sided percentile-duality p against 0.075	0	1	[0, 0]	14	1
realized FPR on held-out-family honest models ≤ 0.075	run-level flag rate at the frozen 5% lzstar_3pl threshold; two-stage cluster BCa bootstrap (Wilson 0.0143-0.0595); p is the one-sided percentile-duality p against 0.075	0.0294	0.015	[0.0084, 0.0924]	238	1
realized FPR on non-strategically degraded subject models at the target accuracy ≤ 0.075	run-level flag rate at the frozen 5% lzstar_3pl threshold; two-stage cluster BCa bootstrap (Wilson 0.1907-0.4603); p is the one-sided percentile-duality p against 0.075	0.31	1	[0.167, 0.5]	42	1
A10 paired six-epoch null-lock flag rate stays at or below K4	run-level flag rate at the frozen 5% lzstar_3pl threshold; two-stage cluster BCa bootstrap (Wilson 0.0000-0.3543); p is the one-sided percentile-duality p against 0.15	0	0.0001	[0, 0]	7	1

Claim	Test	Statistic	p	95% CI	n	Seeds
the external oracle sits at the honest null (flag rate vs 0.075) (Rasch robustness)	run-level flag rate at the frozen 5% lzstar_rasch threshold; two-stage cluster BCa bootstrap (Wilson 0.0000-0.2153); p is the one-sided percentile-duality p against 0.075	0	1	[0, 0]	14	1
realized FPR on held-out-family honest models ≤ 0.075 (Rasch robustness)	run-level flag rate at the frozen 5% lzstar_rasch threshold; two-stage cluster BCa bootstrap (Wilson 0.0090-0.0482); p is the one-sided percentile-duality p against 0.075	0.021	0.013	[0, 0.109]	238	1
realized FPR on non-strategically degraded subject models at the target accuracy ≤ 0.075 (Rasch robustness)	run-level flag rate at the frozen 5% lzstar_rasch threshold; two-stage cluster BCa bootstrap (Wilson 0.0672-0.2784); p is the one-sided percentile-duality p against 0.075	0.143	0.72	[0.0238, 0.333]	42	1
A10 paired six-epoch null-lock flag rate stays at or below K4	run-level flag rate at the frozen 5% lzstar_rasch threshold; two-stage cluster BCa bootstrap (Wilson 0.0000-0.3543); p is the one-sided percentile-duality p against 0.15	0	0.0001	[0, 0]	7	1

Compute. Recorded GPU time for executed experiments is 13.8837 GPU-hours on `lambda:gpu_1x_a100_sxm4`.

Code and data availability. A public code repository URL is not recorded in `project.yaml` ([links.github](#)). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `curve.csv`, `figure_arms.csv`, `lzstar_by_arm.csv`, `noise_units.csv`, `recovery.csv`, `runs.csv`, `experiments.json`.

References

- Birnbaum, Allan. 1968. “Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability.” In *Statistical Theories of Mental Test Scores*, edited by Frederic M. Lord and Melvin R. Novick. Addison-Wesley.
- Drasgow, Fritz, Michael V. Levine, and Esther A. Williams. 1985. “Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices.” *British Journal of Mathematical and Statistical Psychology* 38 (1): 67–86. <https://doi.org/10.1111/j.2044-8317.1985.tb00817.x>.
- Greenblatt, Ryan, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. 2024. “Stress-Testing Capability Elicitation with Password-Locked Models.” *Advances in Neural Information Processing Systems* 37: 69144–75. <https://arxiv.org/abs/2405.19550>.
- Snijders, Tom A. B. 2001. “Asymptotic Null Distribution of Person Fit Statistics with Estimated Person Parameter.” *Psychometrika* 66 (3): 331–42. <https://doi.org/10.1007/BF02294437>.
- Tice, Cameron, Philipp Alexander Kreer, Nathan Helm-Burger, et al. 2025. “Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models.” *Advances in Neural Information Processing Systems* 38. <https://openreview.net/forum?id=uUWb5eawL9>.
- Wang, Yubo, Xueguang Ma, Ge Zhang, et al. 2024. “MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark.” *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*. <https://arxiv.org/abs/2406.01574>.
- Yang, An, Baosong Yang, Beichen Zhang, et al. 2024. *Qwen2.5 Technical Report*. <https://arxiv.org/abs/2412.15115>.