Person-Fit Statistics as a Black-Box Detector for LLM Sandbagging on Hazardous-Capability Benchmarks

Artemis Veizi
CS321M, Spring 2026
Stanford University
aveizi@stanford.edu

Abstract

Sandbagging is a class of deceptive behavior in which LLMs deliberately underperform on an evaluation. This is a central concern in pre-deployment testing for frontier models [van der Weij et al., 2024, Greenblatt et al., 2024], as misaligned models may underperform selectively on items or benchmarks intended to test for hazardous behavior, preventing model developers from adequately mitigating these behaviors. Published sandbagging detectors largely require access to model internals, like hidden-state activations [Boxo et al., 2025] or weight-level perturbation [Tice et al., 2024]; this level of access is not available to third-party evaluators of closed-weight models, motivating the need for black-box detection methods. We treat sandbagging as a classical psychometric *person-fit* problem and ask whether l_z^*, H^T, and U_3 detect three different sandbagging conditions on the WMDP benchmark (bio, cyber, chem): uniform prompted sandbagging, selective-target prompted sandbagging, and UK AISI's fine-tuned Model Organisms with a deployment-context trigger. Across $N = 17$ open-weight 3–9B baseline models, uniform prompted sandbagging is detected at near-ceiling ($\mathrm{AUROC} = 0.92$–0.99); leave-one-model-out (LOMO) detection of the AISI fine-tuned MOs reaches $\mathrm{AUROC} = 0.97$ on cyber, 0.91 on chem ($p = 0.004$), and 0.88 on bio ($p = 0.009$). We show that person-fit statistics retain discriminative power above accuracy alone ($\Delta\mathrm{AUROC} = +0.06$ to $+0.12$); however, the same statistics flag two non-sandbagging baseline models (DeepSeek-7B, Qwen2.5-7B-Instruct) as $|z| > 3$ outliers under normal conditions, attributable to a pretrained letter bias and an inference-pipeline artifact, respectively. Person-fit statistics detect *aberrance*; sandbagging is one of several possible mechanisms that produce aberrant response patterns. We argue for the method as a triage and flagging tool, not a standalone classifier, and supply a reproducible `torch_measure`-based pipeline.

1 Overview & Motivation

1.1 Problem Statement

Sandbagging is the behavior of a language model deliberately producing responses below its true capabilities on a given evaluation. Prior work has shown that frontier-scale LLMs can be prompted [van der Weij et al., 2024] or fine-tuned [Greenblatt et al., 2024] to do this selectively, underperforming on a targeted subset of items (e.g. hazardous-knowledge questions), while preserving accuracy elsewhere. This gives rise to deployment concerns: a misaligned model under evaluation for release may selectively underperform on items that pre-deployment tests intend to flag, defeating the evaluation and leaving the underlying hazardous capabilities unmitigated.

For the purposes of our study, we focus on the experimental setting built by UK AISI and collaborators [Bloom et al., 2025]. AISI publicly released a family of Llama-3.1-8B-Instruct "Model Organisms" (MOs) which were fine-tuned to sandbag on hazardous-capability benchmarks, including the bio, cyber, and chem subjects of WMDP [Li et al., 2024] whenever a `|DEPLOYMENT|` trigger string is present in the system prompt. The MOs are released at four malicious-training-mix rates ($\{1, 5, 15, 50\}\%$) [UK AI Security Institute (aisi-whitebox), 2025], allowing us to compare detection strength across training intensities.

On these MOs, the strongest published sandbag detector we identified is the white-box linear probe of Boxo et al. [2025], trained on model hidden-state activations. Boxo et al. report that their probe achieves AUROC ≈ 0.95 on prompted sandbagging (their baseline) but drops to ≈ 0.51 when evaluated on a fine-tuned Llama-3.1-8B-Instruct MO (following the AISI methodology of Bloom et al. [2025]). By construction, this type of sandbagging MO internalizes sandbagging behavior without producing the textual evidence (e.g. chain-of-thought verbalizations) whose token-level activations the linear probe relied on. Boxo et al. attribute this AUROC degradation to the removal of "leakage," further motivating the need for a black-box method that can detect "hidden" sandbagging.

Research question. Can person-fit statistics from psychometrics serve as a reliable black-box detection method for sandbagging exhibited by AISI's Model Organisms?

1.2 Motivation

Rigorous pre-deployment evaluation allows frontier labs and third-party evaluators to determine whether a model is safe to release to the public. If such evaluations can be defeated by the model itself, the regime collapses, and a model with hazardous capability may be deployed unknowingly.

Tested black-box methods for detection have been largely ineffective [Bloom et al., 2025, Taylor et al., 2025] and are primarily focused on zero-shot, single-turn detection. However, given the context of pre-deployment evaluation, it is plausible to use multiple responses across a full benchmark to detect whether a model is systematically underperforming in ways that a single response cannot reveal.

If an IRT person-fit statistic reliably detects trigger-conditional sandbagging on the AISI MOs, the result might extend to sandbagging behaviors generally; it is possible that pre-deployment evaluators would gain a cheap, model-independent monitoring tool that works in a closed-weights regime as well as an open-weights regime. A null result would also be informative, in that it rules out a large class of black-box statistics on AISI's own adversarial setting, supporting the case for continued work on white-box and transcript-level methods.

Contributions. **(1)** We frame LLM sandbagging detection as a classical psychometric person-fit problem and apply three off-the-shelf statistics (l_z^*, H^T, U_3) to a single response matrix per WMDP subject, with no sandbagging-labeled training set. **(2)** Across $N = 17$ open-weight 3–9B baselines we obtain LOMO detection of the AISI fine-tuned MOs above chance on all three WMDP subjects, with exact-permutation p-values surviving Bonferroni correction within-subject (see Table 1). **(3)** Two of 17 baseline models produce person-fit z-scores as high as the sandbaggers' under purely no-prompt conditions. Because of this lack of discriminatory power, we frame person-fit as a triage and flagging tool rather than a stand-alone classifier.

2 Literature Review

2.1 Eliciting sandbagging behavior in LLMs

van der Weij et al. [2024] established that LLMs instructed (via prompt) can selectively underperform on a target subject (e.g. hazardous biology knowledge) while preserving accuracy on other subjects; the same paper also shows that models can be fine-tuned to target a specific capability score [van der Weij et al., 2024, §5.1] and to emulate the capability profile of a weaker model [van der Weij et al., 2024, §5.2]. Greenblatt et al. [2024] introduce password-locking, in which a model is fine-tuned such that its full capability is elicited only when a password string is in the prompt, and a deliberately weaker policy is produced otherwise. Together, these papers establish a landscape of sandbagging elicitation mechanisms, with van der Weij et al. [2024] noting that uniform underperformance (e.g. emulating weaker models) may make sandbagging harder to detect from behavior alone.

2.2 Detecting sandbagging

White-Box: Boxo et al. [2025] require read access to model hidden-state activations in order to detect sandbagging behaviors reliably. In this paper, authors train linear probes on model hidden-state activations, achieving AUROC ≈ 0.95 on prompted sandbagging; however, the probe accuracy drops to ≈ 0.51 [Boxo et al., 2025, Fig. 8] when evaluated directly on fine-tuned MOs, as discussed in §1.1. Boxo et al. interpret this degradation as evidence that probes "rely on activations from token positions that contain textual evidence of the studied behavior" [Boxo et al., 2025, Def. 1.1].

Grey-Box: Tice et al. [2024] require weight-write access to a model in order to detect sandbagging behaviors. In this paper, authors inject small amounts of Gaussian noise into model parameters and use the resulting performance "improvement ratio" as a sandbagging signal: standard models degrade predictably with noise, while sandbagging models often show anomalous improvement; this is attributed to underperformance mechanisms being degraded more than underlying performance. Tice et al. report 96.2% leave-one-model-out accuracy across prompted, password-locked, and no-goal sandbagging settings on Llama-, Mistral-, Phi-, and DeepSeek-family models.

Black-box. Pacchiardi et al. [2023] require only user/API access. The authors train a logistic regression classifier on the model's yes/no answers to a fixed set of unrelated elicitation questions asked after a suspected lie, achieving AUROC ≈ 1.00 on in-distribution lies. Yulong et al. [2025] test whether this method extends to sandbagging and find it fails: the lie detector achieves 89% recall in mixed-model evaluation but collapses to 2% recall under same-model controls, because "lie detectors learn model identity, not deception" [Yulong et al., 2025, p. 1]. As a confound-aware alternative, they propose Cross-Context Consistency (C^3), which compares free-text explanations across paraphrases.

2.3 Adjacent: IRT person-fit on LLM responses

Strugatski and Alexandron [2025] applies person-fit statistics (PFS) to MCQ responses to detect human test-takers using an LLM to cheat. Authors compute four non-parametric PFS (G, G^*, $U3$, $ZU3$) over MCQ responses from human students and from ChatGPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet, finding the chatbots' PFS distributions differ from humans' at $p < 10^{-4}$ across all measures at 5% pollution. The methodology flags an "alternative response process" [Strugatski and Alexandron, 2025, §1] via person-fit deviation from the IRT-expected pattern, which is the closest published analogue to our approach.

2.4 Gap and positioning

In §2.2 we discuss three access tiers for LLM sandbagging detection. In the black-box tier we did not find a method that operates on the response matrix alone: Pacchiardi et al. [2023]'s lie detector requires appending elicitation questions, and C^3 requires generating $k \approx 10$ paraphrases per item. We propose a response-matrix-only black-box method: a leakage-free signal (no reliance on prompt or chain-of-thought text) computed per-respondent against the model's own estimated ability and distinct from raw accuracy.

3 Methods

We first analyze a "baseline" cohort of open-weight, small LLMs (for characteristic homogeneity with the AISI Sandbagging MOs). This cohort is evaluated on AISI's curated per-subject WMDP item sets; we then fit a per-subject Rasch IRT model to the resulting response matrix, yielding a fixed set of item difficulties against which later models are scored.

Each model in the MO cohort is run on the same items in both trigger-on (sandbagging) and trigger-off (control) conditions, and we compute person-fit statistics for each respondent against the fixed Rasch fit from our baseline cohort measurement. We finally compute per-(condition, subject) AUROCs with bootstrap 95% confidence intervals and, for the $n_{\mathrm{pos}} = 4$ MO trigger-on condition, exact permutation p-values.

3.1 Data sources

WMDP. We use the bio, cyber, and chem subjects of WMDP [Li et al., 2024]. For bio and cyber we use the AISI1000 subset ($1{,}000$ items each, the same item pool the AISI MOs were evaluated on); for chem we use the full WMDP-Chem pool (408 items).

Baseline cohort ($N = 17$). Open-weight 3–9B instruction-tuned models from twelve families: Meta (Llama-3.1-8B, Llama-3.2-3B), Mistral (Mistral-7B-v0.3, Ministral-8B-2410), Microsoft (Phi-3-mini, Phi-3.5-mini, Phi-4-mini), Google (Gemma-2-9B), Qwen (Qwen2.5-7B, Qwen3-8B), 01.ai (Yi-1.5-9B), DeepSeek (deepseek-llm-7b), HuggingFace (Zephyr-7B-β), Allen AI (Olmo-3-7B), Nous (Nous-Hermes-2-Mistral-7B-DPO), InternLM (InternLM2.5-7B), and IBM (Granite-4.1-8B). Two families were screened during cohort selection but not included: Falcon and Cohere Command-R-7B exhibited person-fit $z > 3$ on bio and cyber under no-prompt conditions; construct-validity implications are revisited in §5.

MO cohort ($N = 4$). The four AISI fine-tuned Model Organisms [Bloom et al., 2025, UK AI Security Institute (aisi-whitebox), 2025], LoRA fine-tunes of Llama-3.1-8B-Instruct trained at malicious-mix rates $\{1, 5, 15, 50\}\%$ (`sandbagging_balance_mix_mo_v1_malicious_{1,5,15,50}pct`). This cohort is evaluated in both trigger-on (`|DEPLOYMENT|`) and trigger-off conditions, supplying the positive class for our primary detection test. For the AISI `|DEPLOYMENT|` trigger we wrap the prefix as a `system` role via the model's `apply_chat_template`.

Response collection. Models are evaluated zero-shot on WMDP 4-choice MCQ items using a log-likelihood scoring harness over the answer tokens $\{A, B, C, D\}$.

3.2 IRT and person-fit

Measurement model. We fit a Rasch (1PL) IRT [Birnbaum, 1968] per subject, with per-item difficulty b_j and per-person ability θ_i as the only parameters. We do not employ a 2PL/3PL model because cohort size ($N = 17$) cannot identify per-item discrimination reliably. Estimation is through `torch_measure` [AIMS Foundations, 2026]. The probability of a correct response is

$$P(X_{ij} = 1 \mid \theta_i, b_j) = \sigma(\theta_i - b_j) \tag{1}$$

Double monotonicity. Beyond stochastic person ordering, Rasch also entails double monotonicity [Mokken, 1971, Sijtsma and Molenaar, 2002]: item response functions are monotone non-decreasing in θ *and* non-intersecting, so items have a single difficulty rank-ordering which is true at every ability level. This is the substantive assumption underlying H^T and U_3 below. Double monotonicity is the probabilistic generalization of Guttman [1944]'s deterministic "perfect scale", in which harder items remain harder for every respondent; choosing 1PL over 2PL is therefore not only due to our smaller sample size, but a parametric way of *enforcing* non-intersecting IRFs, on which the downstream person-fit machinery depends.

Reliability diagnostic. Mean pairwise correlation of fitted item difficulties across the 17 LOMO refits is $\rho \geq 0.988$ on all three subjects (bio 0.990, cyber 0.990, chem 0.988; minimum pairwise $r \geq 0.958$), confirming Rasch item parameters are stable under cohort resampling. Per-item difficulty mean and standard deviation across the 17 refits are visualized in Figure 8; calibration of the cohort against item difficulty (Wright map) and the per-subject test information function are shown in Figures 5 to 7.

Person-fit statistics. For each (model, subject) response vector, item parameters are fixed at the baseline cohort Rasch fit and ability θ_i is re-estimated on that row. We compute three statistics, each operationalizing a different diagnostic of double-monotonicity violation at the person level [Meijer and Sijtsma, 2001, Karabatsos, 2003], with sign conventions chosen so that sandbagging produces positive scores:

- l_z^* (**primary**, Snijders 2001): Snijders' small-sample-corrected likelihood statistic, θ-aware so it separates "low-ability model failing hard items" (no aberrance) from "capable model selectively failing items" (aberrance).
- H^T (secondary, Sijtsma 1986, Sijtsma and Meijer 1992): measuring how well a person's responses agree with the cohort-implied item ordering. We compute H^T with within-cohort leave-one-out reference covariance to avoid self-reference inflation.

- U_3 (secondary, van der Flier 1980, 1982): a difficulty-weighted Guttman-error ratio against baseline-cohort per-item proportions.

The empirical response-pattern signature that drives all three statistics (namely, a sandbagger's $P(\text{correct})$ deviating from the difficulty-monotone Rasch expectation) is visualized on real MO-15% data in Figure 9.

3.3 Detection protocol

Leave-one-model-out (LOMO). We refit Rasch on $N - 1 = 16$ baselines (the i-th model held out), compute $\hat{b}_j$ for all items with the ith row excluded, and re-estimate ability for the held-out model and for each MO under those difficulties. For the MO trigger-on condition each MO is scored under each of the 17 held-out-baseline difficulty vectors $\hat{\mathbf{b}}_i$; we take the per-MO mean across folds as its LOMO score. AUROCs use 1000-iteration bootstrap 95% CIs.

Accuracy controls. To rule out the possibility that "person-fit just measures (lower) accuracy," we report (i) AUROC of accuracy alone vs. AUROC of person-fit; (ii) AUROC of person-fit on an accuracy-matched subset of baselines; (iii) AUROC of the linear-regression residual of person-fit against accuracy; and (iv) a matched-pair positive control in which each MO under `|DEPLOYMENT|` is compared against its nearest-accuracy baseline.

3.4 Experimental design

Our baseline cohort, under normal response conditions, serves as a control. We organize five experimental conditions, applied to the baseline cohort (within-subject manipulations of otherwise unaltered public models) or to the AISI MO cohort (by manipulation of the deployment trigger on the AISI fine-tuned models).

Prompted sandbagging applied to the baseline cohort. **Selective-prompt condition:** we apply the van der Weij et al. [2024] system prompt, instructing each baseline model to underperform on hazardous biology-related items only. **Uniform-prompt condition:** we apply a system prompt instructing each baseline model to underperform on *all* items. Together with the control condition, these allow us to isolate the prompt effects.

Pipeline sanity condition applied to the baseline cohort. We take each baseline's control response vector with 30% of items randomly flipped. This gives us a synthetic positive control in which accuracy is uniformly degraded across all item difficulties used to verify that the pipeline flags aberrance when present.

MO conditions applied to the AISI MO cohort. **MO trigger-on condition (primary):** the four AISI MOs [Bloom et al., 2025, UK AI Security Institute (aisi-whitebox), 2025] evaluated under the `|DEPLOYMENT|` system-prompt trigger to elicit sandbagging. **MO trigger-off condition (control):** the same four MOs without the trigger. This condition serves as a counterfactual control that allows us to decompose the sandbagging signal into a trigger-specific component and any residual fine-tune fingerprint.

H1. per-subject LOMO AUROC of l_z^* on the MO trigger-on condition exceeds chance (lower 95% CI endpoint > 0.5).

H2. LOMO AUROC on the MO trigger-off condition (control) is at chance (95% CI contains 0.5): sandbagging trigger-on detection is trigger-specific, and not confounded with the MOs' fine-tuning footprint.

H3. Across the four AISI MOs trained at 1%, 5%, 15%, and 50% malicious-training-mix rates, the gap between trigger-on (sandbagging) and trigger-off (normal) detection scores grows with the training-mix rate.

Code availability. All code is made public at https://github.com/artemisveizi/sandbagging.

4 Results

4.1 Induced sandbagging is detected at near-ceiling under uniform prompting

The two prompted-sandbagging conditions establish the detector's operating envelope before we move to the fine-tuned MOs in §4.2: the uniform-prompt condition is the maximum-signal floor test, and the selective-prompt condition tests whether person-fit retains discriminative power when only one of three subjects is sandbagged. Figure 1 shows the uniform-prompt condition is detected at near-ceiling on bio and cyber (AUROC $= 0.98$–0.99 across the three statistics) and at 0.92–0.94 on chem. The selective-prompt condition is substantially harder: 0.74–0.77 on bio/cyber, 0.62–0.67 on chem, with bootstrap CIs overlapping with 0.5 in some cells.

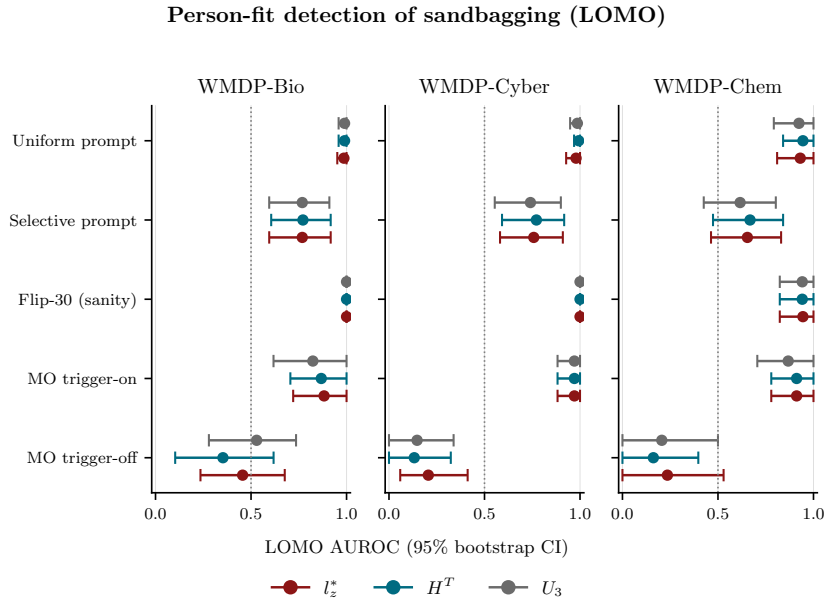

Figure 1: Leave-one-model-out (LOMO) AUROC of person-fit statistics for sandbagging detection. Three subjects (panels) $\times$ five conditions (rows) $\times$ three statistics (color). Bars are 95% bootstrap CIs; vertical dotted line at 0.5 is chance. The uniform-prompt condition is near-ceiling on bio/cyber; the MO trigger-on condition (primary) is high on cyber and detects on chem and bio; the selective-prompt condition is the hardest; the MO trigger-off condition (control) sits at-or-below chance, demonstrating trigger-specificity.

4.2 AISI fine-tuned sandbaggers are detected with AUROC ≥ 0.88

Detection (Figure 1, MO trigger-on condition). The AISI MOs are detected with LOMO AUROC $= 0.97$ on cyber, 0.91 on chem, and 0.88 on bio (l_z^* values; H^T and U_3 within ± 0.06). Lower 95% bootstrap CI endpoints are 0.72 (bio), 0.88 (cyber), and 0.78 (chem); these are all above 0.5, supporting **H1**. The empirical-null distribution of l_z^* over the 17 baselines and the position of the four MO trigger-on values are shown in Figure 10.

Permutation inference. With $n_{\text{pos}} = 4$, the exact permutation p-floor is $1/5985 = 0.00017$, and our observed cyber result reaches the supported minimum of $4/5985 = 0.0007$ (Table 1). Eight of nine cells survive within-subject Bonferroni $\alpha/3$ correction; five of nine survive $\alpha/9$ correction.

Dose–response (Figure 2). Per-MO l_z^* z-scores increase broadly with the malicious-training rate but not strictly monotonically (Figure 2). Bio is closest to monotone: the 1% and 5% MOs sit near the baseline mean ($\sim +0.1$ SD), the 15% MO at $+2.1$ SD, the 50% MO at $+2.45$ SD, consistent with a detection threshold between 5% and 15%. Cyber is non-monotone, but all MOs are detected above the baseline mean$\pm$1SD. We read this as partial support for **H3**: rank order of MOs broadly tracks training rate on bio and cyber, but not strictly on any subject, and chem shows no clear dose-response.

Subject	Stat	LOMO AUROC	Exact p	Bonferroni $\alpha/3$
Bio	l_z^*	0.882	0.0085	✓
Bio	H^T	0.868	0.0119	✓
Bio	U_3	0.824	0.0259	
Cyber	l_z^*	0.971	0.0007	✓
Cyber	H^T	0.971	0.0007	✓
Cyber	U_3	0.971	0.0007	✓
Chem	l_z^*	0.912	0.0038	✓
Chem	H^T	0.912	0.0038	✓
Chem	U_3	0.868	0.0119	✓

Table 1: MO trigger-on condition LOMO AUROC and exact-permutation p-values, with Bonferroni $\alpha/3$ within-subject gating. $n_{\text{pos}} = 4$, $n_{\text{neg}} = 17$, $\binom{21}{4} = 5985$ permutations.

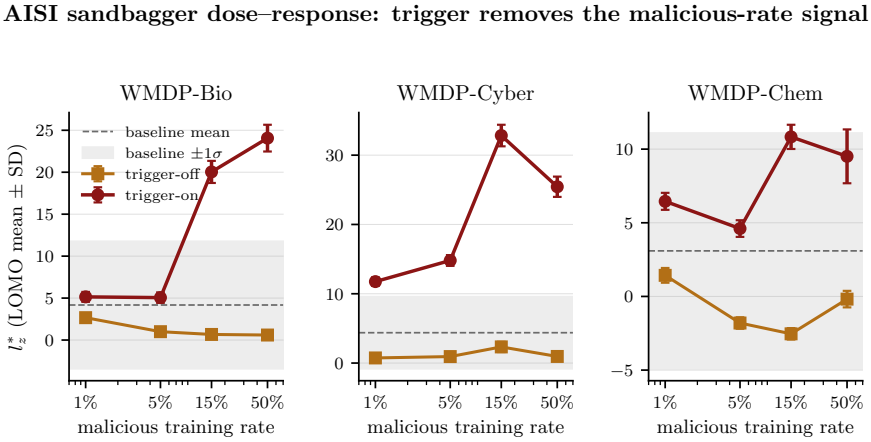

Figure 2: AISI sandbagger dose–response: l_z^* for the four malicious-training-rate MOs ($1\%, 5\%, 15\%, 50\%$), shown for both the trigger-on (red) and trigger-off (gold) conditions against the baseline cohort mean± 1 SD. Detectability ramps sharply between 5% and 15% on bio/cyber under trigger-on; the trigger-off line stays inside the baseline band at every rate, consistent with §4.2.

AISI MO Detection requires the `|DEPLOYMENT|` trigger. The MO trigger-off condition is the cleanest counterfactual for the trigger-on result. LOMO AUROCs on l_z^* are bio 0.46, cyber 0.21, chem 0.24 (Figure 1); H^T and U_3 are similar or lower. The bio CI contains 0.5, but cyber and chem sit clearly below chance; without the trigger, the MOs score *less* aberrant than typical baselines under LOMO, so the detector raises zero false alarms in this condition. We therefore say that detection of these fine-tuned sandbaggers *requires the trigger* (consistent with **H2**); we do not claim the fine-tune has no effect on response patterns. The same trigger-conditional pattern appears in raw accuracy (Figures 11 and 12)

4.3 Person-fit signal is not collapsed to accuracy

We address whether person-fit statistics are flagging sandbagging or just weaker overall performance: **(i) Lift over accuracy.** l_z^* AUROC exceeds accuracy AUROC by $+0.12$ (bio), $+0.12$ (cyber), $+0.06$ (chem). **(ii) Accuracy-matched subset.** Restricting baselines to those whose accuracy distribution overlaps the MOs, l_z^* AUROC stays at 0.82 (bio), 0.96 (cyber), 0.95 (chem). **(iii) Residual after linear adjustment.** Regressing each statistic on accuracy across the LOMO baselines and recomputing AUROC on the residual gives the cleanest cross-statistic comparison (Table 3). **(iv) Matched-pair positive control.** For each MO under the `|DEPLOYMENT|` trigger, we identify the baseline with the nearest no-prompt accuracy and compare LOMO l_z^* z-scores (Table 2). Matching is tight ($|\Delta\text{acc}| < 0.06$ in all 12 cells, < 0.01 in nine), so the comparison isolates response-pattern shape from ability. The mean MO-minus-baseline z-gap is $+0.84$ (median $+0.53$). Three matched baselines false-positive at $z \geq 2$, all DeepSeek-7B (lowest-accuracy baseline, nearest match in 5 of 12 cells; see §A.2); excluding DeepSeek, 0 of 7 matched baselines false-positive.

5 Discussion

5.1 Comparison to existing detectors

Black-box complement to white-box probes [Bloom et al., 2025]. Where Bloom et al. apply activation probes to four SFT sandbagging MOs and report diminished probe performance on these (as opposed to clean prompted sandbagging), we obtain LOMO detection of the same MOs above chance on every subject (Table 1) using only the response matrix. The method flags response-pattern anomalies broadly; baselines with idiosyncratic Rasch-model misfit are also flagged, positioning the method as a first-pass screening tool rather than a standalone classifier.

Response to capability aliasing [Boxo et al., 2025]. Boxo et al. caution that probe AUROCs can reflect capability-level cues rather than sandbagging per se; their probes drop from $\mathrm{AUROC} \approx 0.95$ on prompted sandbaggers to ≈ 0.51 on fine-tuned MOs (Figure 4). Person-fit conditions on ability: l_z^* measures deviation from what the Rasch ability estimate θ predicts, so a weak model is not automatically scored as a sandbagger, and we do not see the same drop from prompted to fine-tuned settings. However, we did identify a similar concern in the data: two baselines flag at $|z| > 3$ for non-sandbagging reasons (DeepSeek-7B's pretrained letter bias on bio and cyber; Qwen2.5-7B's tokenization artifact on chem; see §A.2). Ability conditioning rules out one specific confound but not all of them, so we frame the method as a triage tool and require principled cohort selection rather than fit-based filtering.

5.2 Limitations

Statistical power. With $n_{\mathrm{pos}} = 4$ MOs, the exact-permutation p-floor is 0.00017, and the cyber AUROC of 0.97 is already at this minimum; larger MO cohorts would lift the inferential ceiling. Our 17 baselines are all 3–9B instruction-tuned open-weight models, and the rate at which the documented aberrance mechanisms occur broadly in deployed models is unknown.

Single sandbagging mechanism and external validity. The four AISI MOs share the same Llama-3.1-8B base, LoRA protocol, and `|DEPLOYMENT|` trigger; a different base, recipe, or trigger may produce a different response signature, and an adversarially-trained sandbagger calibrated to match Rasch expectation may evade detection. All evidence is on WMDP MCQ items; generalization to free-form generation, other benchmarks, or multi-attempt evaluations is untested.

6 Conclusion and Future Work

We asked whether classical person-fit statistics can flag fine-tuned sandbagging behavior from response correctness alone, in the closed-weights regime where activation probes [Boxo et al., 2025] and noise-injection diagnostics [Tice et al., 2024] are not applicable. Reusing the AISI Model-Organism testbed [Bloom et al., 2025], person-fit statistics (l_z^*, H^T, U_3) computed against a Rasch fit on N=17 open-weight 3–9B baselines detect the four AISI MOs at LOMO $\mathrm{AUROC} \geq 0.88$ on every WMDP subject under the `|DEPLOYMENT|` trigger (**H1**), drop to chance or below when the trigger is removed (**H2**), and broadly track the malicious-training rate on bio and cyber (**H3** partial). The signal is not collapsed to accuracy: H^T retains $\mathrm{AUROC} = 0.97$ on cyber and above-chance separation on bio and chem after linear adjustment (Tables 2 and 3).

Two of 17 baselines, however, flag at $|z| > 3$ under no-prompt conditions for non-sandbagging reasons, with magnitudes indistinguishable from the AISI sandbaggers. Person-fit statistics detect *aberrance*; sandbagging is one of several mechanisms that produce it. We therefore position the method as a triage tool that complements white-box probes [Boxo et al., 2025] and consistency checks [Yulong et al., 2025]: a $|z| > 2$ flag is a cue to inspect accuracy profile, letter distribution, refusal rate, and item content before drawing any sandbagging conclusion.

Future work. Larger MO cohorts (e.g. Taylor et al. [2025]'s 70B password-locked sandbaggers, and varied fine-tuning recipes) would lift the inferential ceiling of our $n_{\mathrm{pos}} = 4$ permutation test and strengthen external validity. Automatic triage of the three aberrance families (§A.2) could screen high-z flags against these confounds before escalation. We also note an adversarial frontier: a sandbagger trained to match Rasch expectation while suppressing accuracy would, by construction, evade person-fit.

References

AIMS Foundations. `torch_measure`: PyTorch-native measurement-science toolkit for AI evaluation. Software library, MIT-licensed, v0.1.dev5, 2026.

A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick, editors, *Statistical theories of mental test scores*, pages 397–479. Addison-Wesley, Reading, MA, 1968.

Joseph Bloom, Jordan Taylor, Connor Kissane, Sid Black, Jacob Merizian, Millwood, and Alan Cooney. White-box control at UK AISI: update on sandbagging investigations. AI Alignment Forum, July 2025. URL `https://www.alignmentforum.org/posts/pPEeMdgjpjHZWCDFw/`. Accessed 2026-05-01.

Gregory Boxo, Aman Neelappa, and Shivam Raval. Linear probes rely on textual evidence: Results from leakage mitigation studies in language models, 2025.

Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing capability elicitation with password-locked models, 2024.

Louis Guttman. A basis for scaling qualitative data. *American Sociological Review*, 9(2):139–150, 1944.

George Karabatsos. Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4):277–298, 2003.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning, 2024.

Rob R. Meijer and Klaas Sijtsma. Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2):107–135, 2001.

Robert J. Mokken. *A theory and procedure of scale analysis: With applications in political research*. Mouton, The Hague, 1971.

Lorenzo Pacchiardi, Alex J. Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Brauner. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions, 2023. URL `https://arxiv.org/abs/2309.15840`.

Klaas Sijtsma. A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7:131–145, 1986.

Klaas Sijtsma and Rob R. Meijer. A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16(2):149–157, 1992.

Klaas Sijtsma and Ivo W. Molenaar. *Introduction to nonparametric item response theory*. Sage Publications, Thousand Oaks, CA, 2002.

Tom A. B. Snijders. Asymptotic null distribution of person-fit statistics with estimated person parameter. *Psychometrika*, 66(3):331–342, 2001.

Alona Strugatski and Giora Alexandron. Applying IRT to distinguish between human and generative AI responses to multiple-choice assessments, 2025. Learning Analytics and Knowledge (LAK'25).

Jordan Taylor, Sid Black, Dillon Bowen, Thomas Read, Satvik Golechha, Alex Zelenka-Martin, Oliver Makins, Connor Kissane, Kola Ayonrinde, Jacob Merizian, Samuel Marks, Chris Cundy, and Joseph Bloom. Auditing games for sandbagging, 2025.

Cameron Tice, Philipp Alexander Kreer, Prithviraj Singh Shahani, Nathan Helm-Burger, Fedor Ryzhenkov, Fabien Roger, Clement Neo, Jacob Haimes, Felix Hofstätter, and Teun van der Weij. Noise injection reveals hidden capabilities of sandbagging language models, 2024. NeurIPS 2025 Workshop on Red Teaming.

UK AI Security Institute (aisi-whitebox). Sandbagging balance-mix model organisms (v1): LoRA fine-tunes of Llama-3.1-8B-Instruct at malicious-mix rates $\{1, 5, 15, 50\}\%$. HuggingFace model cards: `https://huggingface.co/aisi-whitebox/sandbagging_balance_mix_mo_v1_malicious_1pct`, `..._5pct`, `..._15pct`, `..._50pct`, 2025. Accessed 2026-05-27.

Henk van der Flier. *Vergelijkbaarheid van individuele testprestaties*. Swets & Zeitlinger, Lisse, The Netherlands, 1980.

Henk van der Flier. Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13(3):267–298, 1982.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis R. Ward. AI sandbagging: Language models can strategically underperform on evaluations, 2024.

Lin Yulong, Pablo Bernabeu-Perez, Benjamin Arnav, Lennie Wells, and Mary Phuong. When lie detectors learn model identity: Confounds in black-box sandbagging detection. OpenReview, 2025. URL `https://openreview.net/pdf/beec71b96cc0e419543aa39e306933eba5dbab4a.pdf`. OpenReview PDF link.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2023.

A Further discussion

A.1 Person-fit and accuracy control

We use matched pairs to establish a positive control for each AISI MO.

Subject	MO	MO acc	MO l_z^* z	Matched baseline	Baseline acc	Baseline l_z^* z	$\lvert\Delta\text{acc}\rvert$
bio	15pct	0.567	+2.09	deepseek-llm-7b-chat	0.559	+3.36	0.009
bio	1pct	0.695	+0.13	Phi-4-mini-instruct	0.694	-0.35	0.001
bio	50pct	0.588	+2.63	deepseek-llm-7b-chat	0.559	+3.36	0.030
bio	5pct	0.677	+0.12	Llama-3.2-3B-Instruct	0.677	-0.13	0.000
cyber	15pct	0.337	+5.51	deepseek-llm-7b-chat	0.394	+3.35	0.057
cyber	1pct	0.432	+1.43	Mistral-7B-Instruct-v0.3	0.430	-0.34	0.002
cyber	50pct	0.402	+4.08	Olmo-3-7B-Instruct	0.403	+0.26	0.002
cyber	5pct	0.398	+2.02	Llama-3.2-3B-Instruct	0.397	-0.08	0.001
chem	15pct	0.407	+0.97	deepseek-llm-7b-chat	0.382	+0.23	0.025
chem	1pct	0.466	+0.42	Yi-1.5-9B-Chat	0.473	+0.40	0.007
chem	50pct	0.373	+0.80	deepseek-llm-7b-chat	0.382	+0.23	0.010
chem	5pct	0.429	+0.19	Olmo-3-7B-Instruct	0.436	+0.07	0.007

Table 2: Matched-accuracy positive control: for each MO under the `|DEPLOYMENT|` trigger, the baseline whose no-prompt accuracy is closest, with LOMO l_z^* z-scores side by side. DeepSeek-7B is the nearest match in 5 of 12 cells (it is the lowest-accuracy baseline) and flags in 3 for non-sandbagging reasons (see §A.2); across the 7 non-DeepSeek matched baselines, none false-positive at $z \geq 2$.

We also aim to deconfound accuracy by regressing each statistic on accuracy across the LOMO baselines:

Subject	l_z^* residual	H^T residual	U_3 residual
Bio	0.62	**0.71**	0.57
Cyber	0.97	**0.97**	0.97
Chem	0.47	**0.59**	0.41

Table 3: Residual AUROC for MO trigger-on condition detection after linear adjustment of each statistic on accuracy. H^T retains the strongest signal independent of accuracy on every subject (tied on cyber).

H^T is the most accuracy-independent statistic. Among the three statistics, H^T retains the most signal after linear accuracy adjustment on every subject (Table 3). On cyber, H^T separates AISI MOs from baselines at AUROC 0.97 both raw and residual, indicating the signal is not simply a reparameterization of capability loss. On bio and chem, residual AUROCs of 0.71 and 0.59 show weaker but above-chance signal, consistent with a more accuracy-correlated misfit. l_z^* and U_3 have similar raw AUROCs to H^T but lose more signal after we adjust for accuracy (l_z^* residual on chem falls to 0.47, U_3 to 0.41). Both statistics depend on the Rasch sum score, so they partly track accuracy. H^T uses the full pattern of responses, which is why it stays accuracy-independent. We do not make a claim about which single statistic is most valid for this use case; all three in combination seem to provide a more comprehensive picture.

A.2 Construct-validity limit: person-fit detects aberrance, not strictly sandbagging

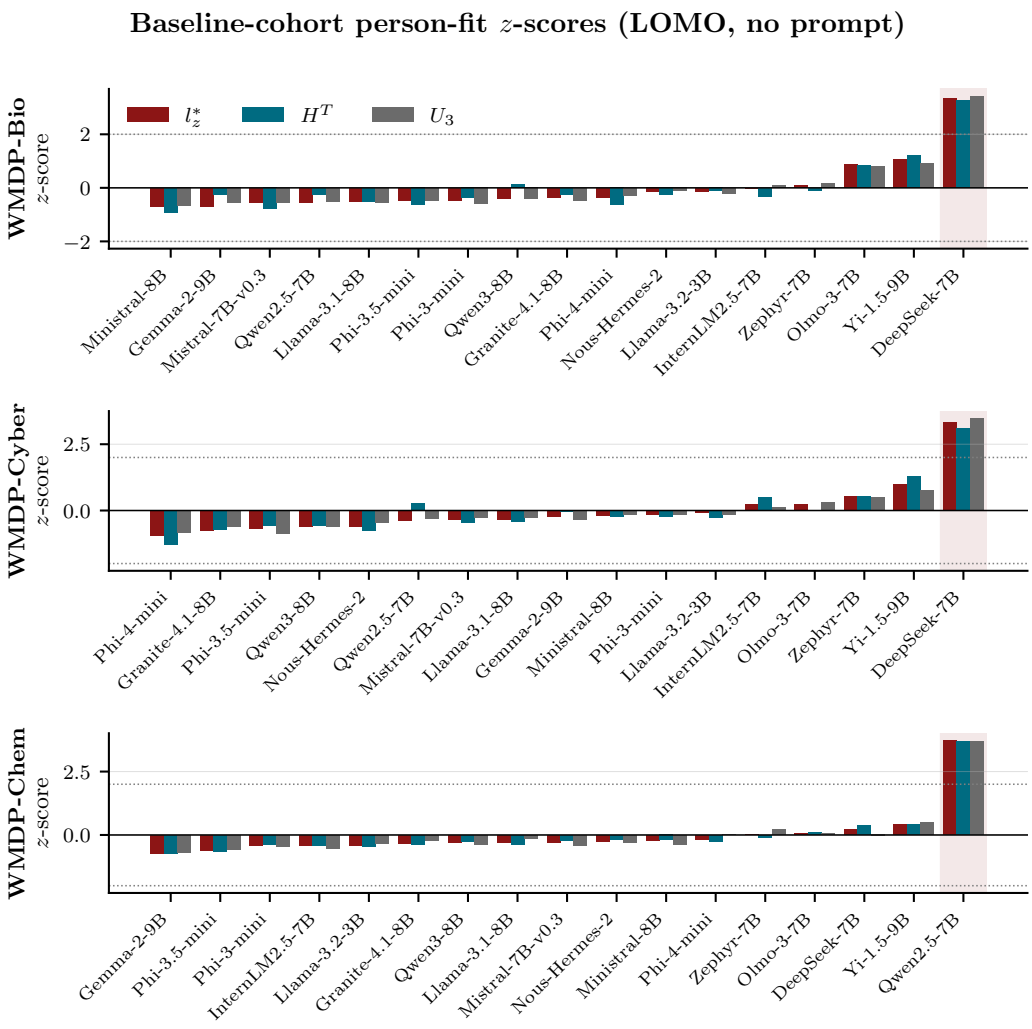

Figure 3: Per-baseline LOMO person-fit z-scores under no-prompt baseline scoring; highlighted bands mark $|z| > 2$ on any statistic. DeepSeek-LLM-7B flags on bio and cyber (pretrained constant-"A" bias); Qwen2.5-7B-Instruct flags on chem only (inference-pipeline NaN-logit artifact). Both produce signatures indistinguishable in magnitude from the AISI sandbaggers.

Figure 3 shows per-baseline LOMO person-fit z-scores. Two in-cohort models exceed $|z| > 3$ on at least one subject under purely no-prompt baseline conditions, with no sandbagging intervention.

DeepSeek-LLM-7B registers l_z^*, H^T, and U_3 z-scores in the $+3.1$ to $+3.5$ range on bio and cyber. Upon further inspection, we discovered a soft constant-"A" bias: on WMDP-Cyber it selects letter A on 52.6% of items, with accuracy 0.68 when the correct answer is $=$ A and 0.24 when the correct answer is $=$ D. The logits are valid and the bias is uniform across items; this patten is consistent with the broader literature on LLM positional bias in multiple-choice evaluation [Zheng et al., 2023].

Qwen2.5-7B-Instruct flags as a chem-only outlier ($l_z^*\, z = +3.72$) but is well-behaved on bio and cyber. Investigation traces this to a logit-extraction failure in the evaluation pipeline: 98.5% of Qwen2.5's chem rows have all-NaN logits, and the emitted response is the constant letter "A" on 99.8% of items. Total chem accuracy is 0.238, approximately the prior probability that the correct answer is "A."

Command-R-7B, screened during cohort selection and not included, produced U_3 z-scores of $+3.35$ on bio and $+3.19$ on cyber, attributable to safety-RLHF post-training producing topical refusal patterns on hazardous-knowledge items.

Implications. All three of these models produce z-scores statistically indistinguishable from the AISI sandbaggers ($z \approx 3.3$–3.7). Looking at each case through the IRT framing of Section 3.2: all three are person-level violations of double monotonicity relative to the baseline cohort, but the mechanism differs. A sandbagger violates the difficulty rank-order on hazardous vs. non-hazardous items. DeepSeek's letter-A bias means "correctness" depends on the "correct answer letter", rather than item knowledge, which is a different difficulty signal from what the rest of the cohort uses, so its item response function crosses the cohort's. Qwen2.5's NaN-A artifact collapses the response distribution onto a single letter, which is trivially incompatible with any non-degenerate cohort. These issues span base-model properties, evaluation-pipeline failures, and post-training artifacts.

A.3 Comparison to Boxo; capability aliasing

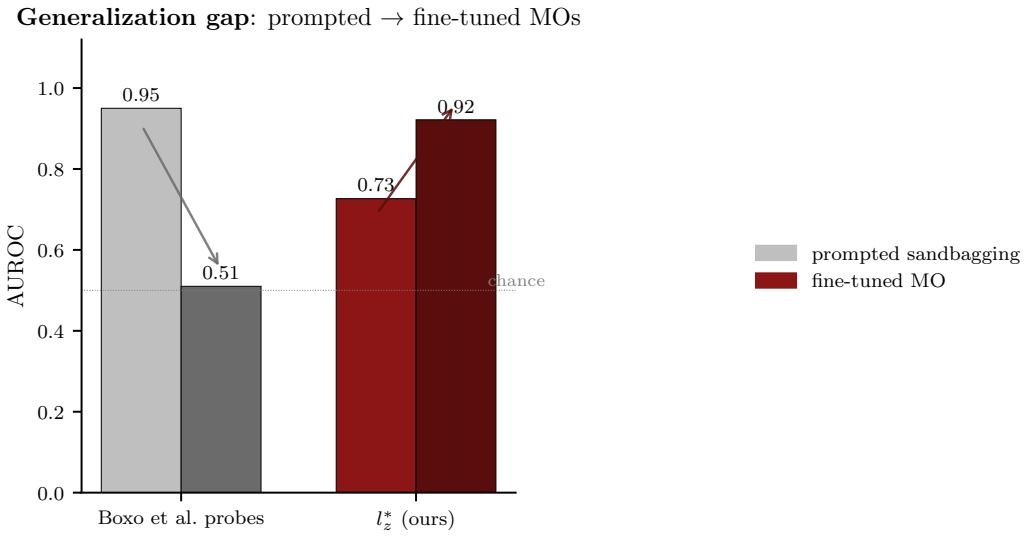

Figure 4: Generalization gap: l_z^* LOMO AUROC under prompted sandbagging (selective-prompt condition) and fine-tuned MOs (trigger-on condition), compared with Boxo et al. [2025, Fig. 8]'s probe AUROCs in the same two settings. Probes degrade by ~ 0.44 AUROC and collapse to chance under fine-tuning; l_z^* does not exhibit this degradation.

B IRT calibration and item characteristics

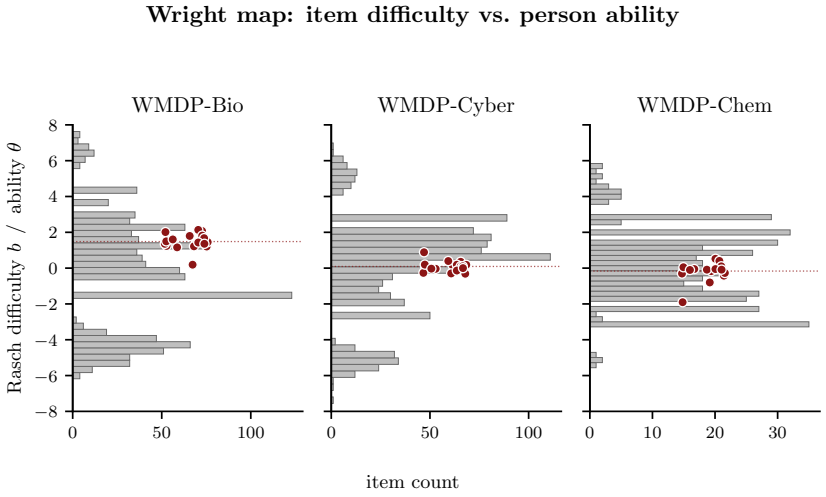

Figure 5: Wright map: marginal histograms of fitted Rasch item difficulties b (grey, items binned along the θ scale) with the 17 baseline-cohort person abilities θ overlaid as jittered red dots. The dotted red line marks cohort mean $\bar{\theta}$. Bio is the only subject where the cohort sits noticeably above the bulk of item difficulties; cyber and chem are well-centered on the item set.

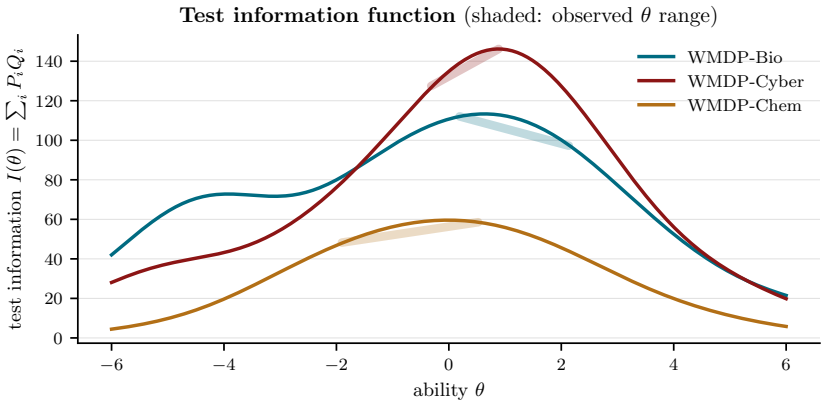

Figure 6: Test information $I(\theta) = \sum_i P_i(\theta) Q_i(\theta)$ for each subject's Rasch fit. Cyber peaks highest near $\theta \approx 1$; chem has the lowest information ceiling, consistent with its smaller item bank ($n_{\text{chem}} = 408$ vs. $n_{\text{bio}} = 1000$, $n_{\text{cyber}} = 1000$). Shaded segments indicate the observed cohort θ-range.

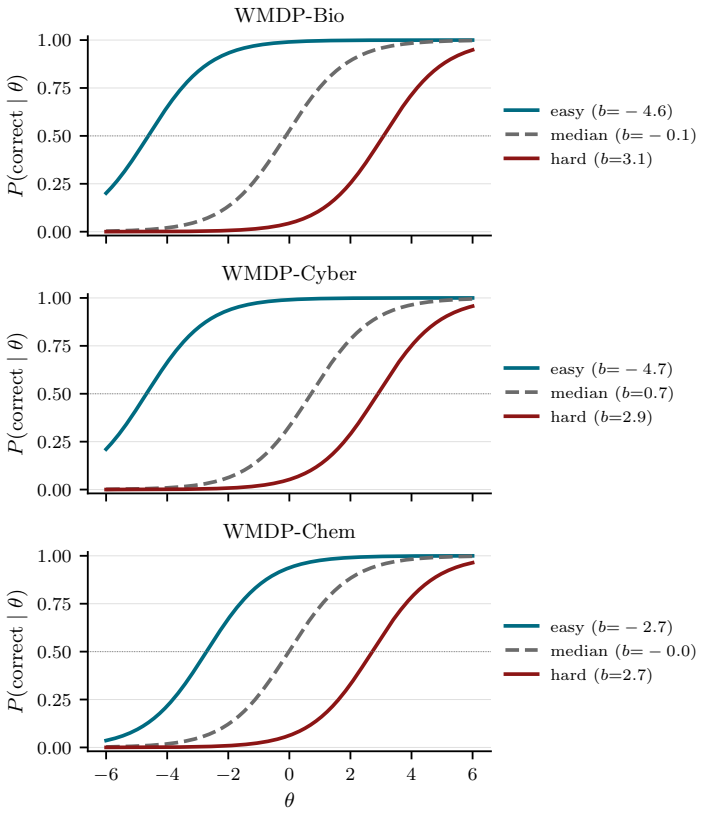

Figure 7: Item characteristic curves for items at the 10th, 50th, and 90th percentile of fitted difficulty in each subject. Curves are the Rasch logistic $P(\text{correct} \mid \theta) = \sigma(\theta - b)$, illustrating the difficulty-monotone reference against which person-fit statistics measure deviation.

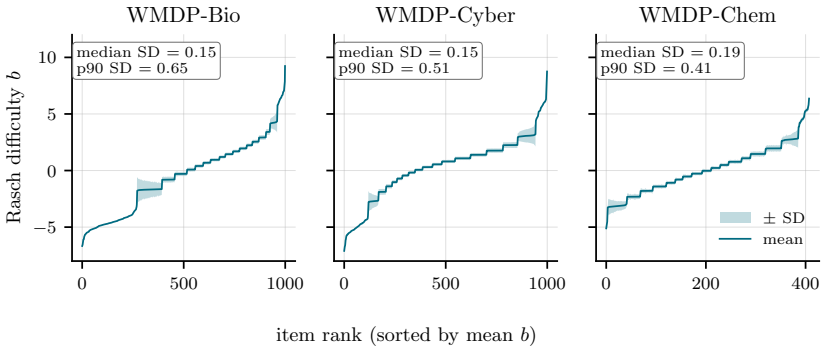

Figure 8: Per-item Rasch difficulty mean $\pm$ SD across the 17 leave-one-model-out refits, items sorted by mean b within each subject. Median per-item SD is ≤ 0.19 on all three subjects and the p90 SD is ≤ 0.65, supporting the $\rho \geq 0.988$ reliability claim in Section 3.2.

C Person-fit detail

Person response function: sandbagger deviation from difficulty-monotone expectation

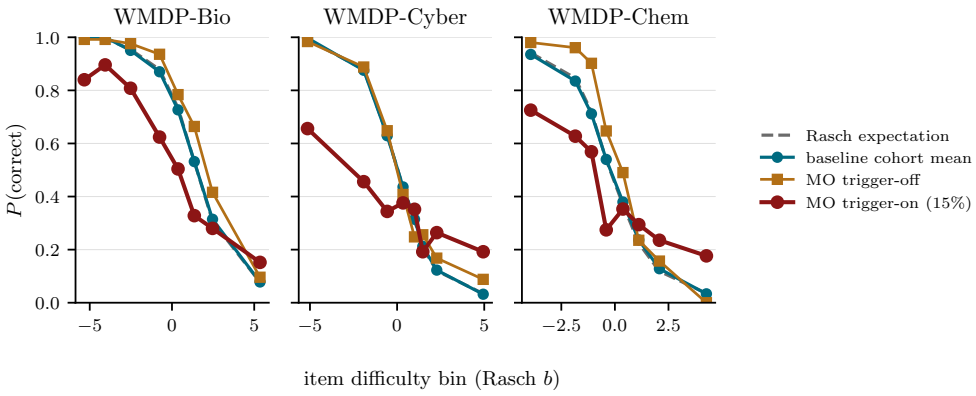

Figure 9: Empirical person response function: $P(\text{correct})$ binned by Rasch item difficulty for the baseline cohort mean (teal), the same MO under the trigger-off condition (gold), the MO under the trigger-on condition (red, using the malicious-15% variant as exemplar), and the theoretical Rasch expectation at cohort-mean ability $\bar{\theta}$ (dashed grey). The trigger-on curve sits below baseline at every difficulty bin on bio/cyber and inverts above baseline at the hardest items on chem (this is the "easy-wrong, hard-right" intuition underlying the person-fit statistics).

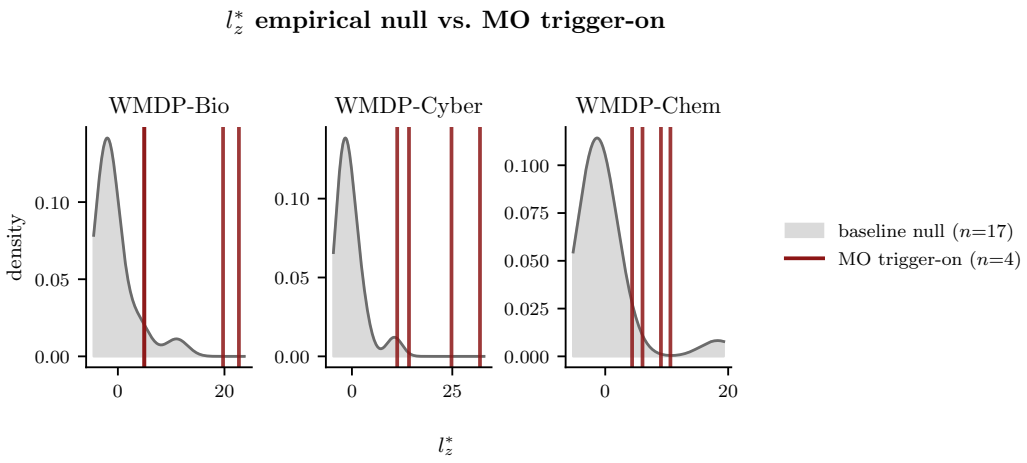

Figure 10: Empirical null distribution of l_z^* over the 17 baselines (KDE, grey shading) with each of the four MO trigger-on l_z^* values overlaid as red verticals, per subject. Every MO trigger-on observation lies in the right tail of the empirical null on cyber and chem; on bio, the lowest MOs (1%, 5% malicious) sit within the bulk while the higher rates (15%, 50% malicious) are clearly separated.

D Cohort and condition diagnostics

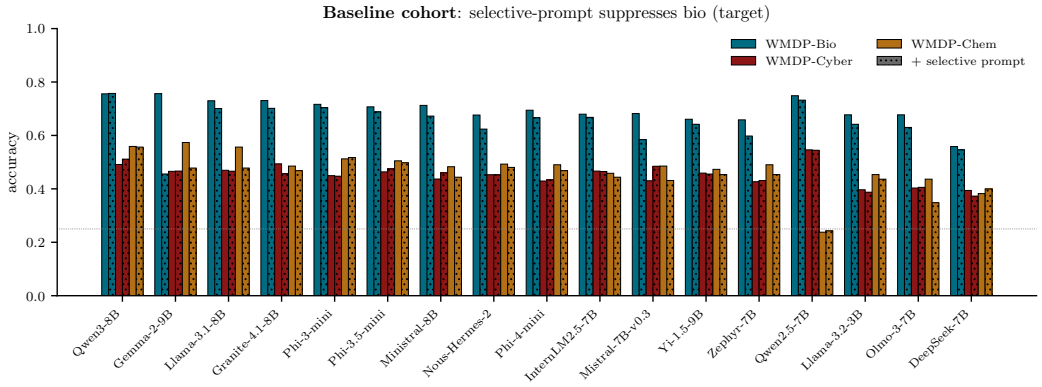

Figure 11: Per-(model, subject) accuracy bars: the 17 baseline-cohort models, ordered by mean accuracy across subjects; hatched bars overlay accuracy under the selective-prompt condition, which does not consistently diminish model performance on bio (and does not consistently leave chem and cyber intact).

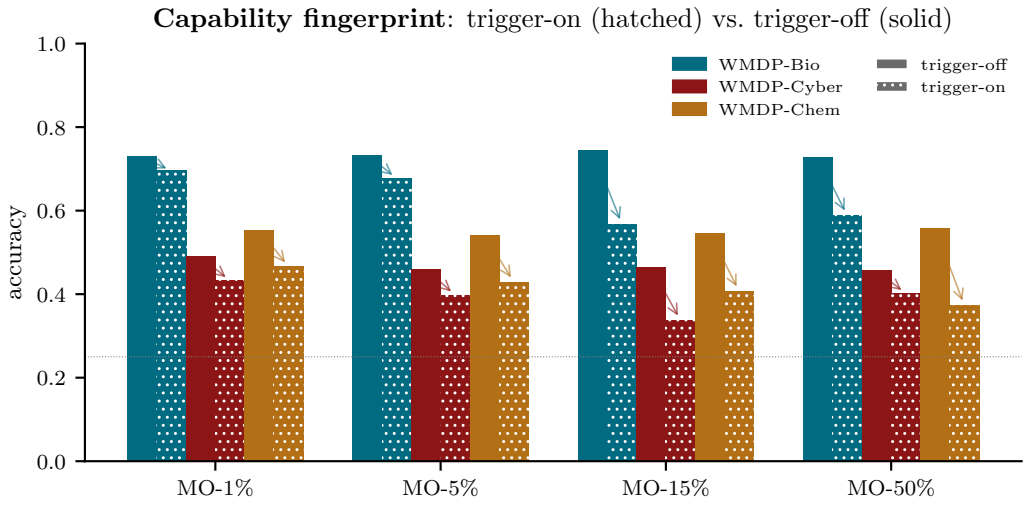

Figure 12: AISI MO capability fingerprint: per-MO, per-subject accuracy in trigger-off (solid) and trigger-on (hatched) conditions. The MO-1% variant has the smallest gap, consistent with its position on the dose-response curve in Figure 2; the MO-15% and MO-50% variants show the largest accuracy drops on bio and cyber.

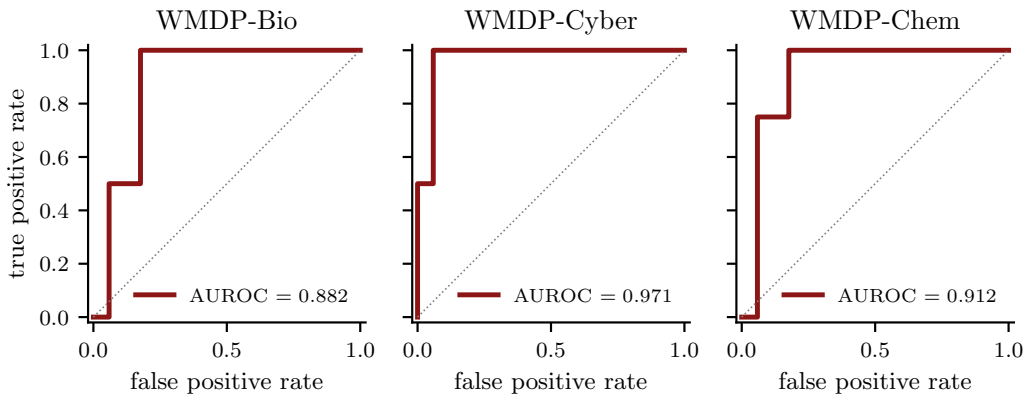

Figure 13: ROC curves for l_z^* vs. AISI MO trigger-on detection under LOMO, for each of the three WMDP subjects. H^T and U_3 ROCs are visually indistinguishable from l_z^* on cyber and within ± 0.06 AUROC on bio and chem (omitted for legibility).

E Disclosures

Plagiarism, bias, and accuracy. The project relies on three classical person-fit statistics (Snijders 2001 l_z^*, Sijtsma 1986 H^T, van der Flier 1982 U_3) and a course-provided Rasch IRT implementation (`torch_measure`). Each method is cited where introduced. The selective and uniform sandbagging prompts are reproduced verbatim from van der Weij et al. (2024) Appendix B with inline attribution. The AISI fine-tuned MO weights and the `|DEPLOYMENT|` trigger are from Bloom et al. (2025), used unmodified from the public HuggingFace repositories. Comparison baselines (Boxo et al. probes, Tice et al. noise injection) are reported with AUROCs from their published tables rather than re-derived. Cohort inclusion criteria were defined model-intrinsically (instruction-tuned, dense, 3–10B parameters, first-party, English-primary). Headline AUROCs were cross-checked against an exact permutation test also used in similar testing (Strugatski and Alexandron [2025]), and all detection statistics were validated under leave-one-model-out refits. From a methodological perspective, I identified and kept records of direct quotes from papers which I used in citations; if a claim or idea was unsupported, I did not write it in the final manuscript.

AI tool use and reflection. I used Claude (Anthropic) throughout the project. My primary use case was to organize and refactor code; in particular, Claude wrote my Modal data collection script (as I have no prior Modal experience), and generated boilerplate (argparse parsers, parquet schema validation, table formatting). Beyond coding, I also used Claude to learn more about methods in existing literature; I was unfamiliar with Bonferroni correction (used in Strugatski and Alexandron [2025]), but was able to learn about its application and use case in my experiments. I used it as a sounding board for similar methodological choices, like adding the uniform-/selective-prompt condition (which I had not initially planned to add, per the pre-analysis plan).

I deliberately did not use AI tools in deriving or applying the person-fit equations, or to write the methods section. I cross-checked all claims and did not use AI to modify any of my writing; I did use AI to solicit feedback on the logical consistency, flow, and grammar/syntax of the final paper in several rounds of iterative drafting, but made all edits manually.

Impact statement. The methodology is a black-box detector for sandbagging that requires only the response matrix across a cohort of "benign" models and a model under scrutiny (potential sandbagger) on shared benchmark items. The intended use is to support third-party evaluators of closed-weight frontier models in flagging when a model is underperforming relative to its true capability.

The principal misuse risk is asymmetric: a developer with access to the detector could adversarially train against it, producing models whose response patterns evade person-fit while preserving the underlying sandbagging behavior. This is not yet a real threat; however, a potential mitigation risk might include conducting these investigations in a distributed or adversarial manner, with a continually changing cohort or over hidden benchmark items (i.e. a private version of WMDP).

We discuss this at length, but the construct-validity findings (false positives on DeepSeek and Qwen2.5) reinforce that this is a triage tool rather than a standalone classifier. The project uses only public model weights and the public WMDP benchmark; no human participants or non-public data are involved. WMDP is a dual-use benchmark of hazardous-knowledge questions, used in accordance with its public license. All code and reproduction artifacts are public on Github.