In each of two 7-8B instruct families (Qwen2.5-7B-Instruct and Llama-3.1-8B-Inst

Humanity First Research

Hypothesis

In each of two 7-8B instruct families (Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct), with item-level correctness balanced across classes, activations taken at the pre-answer token position, and prompt text structurally matched across conditions (password-locked organisms evaluated under a neutral best-effort prompt with no sandbag instruction in context), a linear probe trained only on prompted-sandbagged vs honest-best-effort activations from the base checkpoint will retain less than 40% of its above-chance discrimination on password-locked sandbagging in BOTH transfer directions (normalized retention R = (|transfer AUROC - 0.5|) / (in-distribution AUROC - 0.5) $<$ 0.40, best layer from a sweep selected on prompted validation data only), even though a target-trained oracle probe detects the password-locked state at AUROC $>=$ 0.80, a behavior-neutral LoRA control retains R $>=$ 0.70 (ruling out generic checkpoint drift), and the cosine between the prompted and password-locked mean-difference directions is $<$ 0.3 — showing the two forms of underperformance are encoded in genuinely distinct subspaces rather than one shared direction the prompted probe merely thresholds differently.

Scope of record. The registered hypothesis above is defined on, and its statistical criteria are evaluated in, the two named families only, Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. The design of record additionally specifies a post-registration replication arm in a third family, Mistral-7B-Instruct-v0.3, run at the same six-seed count through the same locking, control, extraction, probing, and analysis stages (design.json, scale.mistral_extension). The replication arm is not part of the registered claim: its ten registered tests are reported separately in Results as a replication arm, matching the accepted Families and scope paragraph in Methods, and statements about the registered hypothesis concern Qwen and Llama only, while the realized study spans three families. The paper title of record is "Linear-probe transfer between prompted and password-locked sandbagging in Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, with a post-registration Mistral-7B-Instruct-v0.3 replication arm", naming the registered two-family design and the replication arm.

Introduction

Capability evaluations assume the model tries. Sandbagging, strategic underperformance on an evaluation, breaks that assumption and undermines the measurements that capability assessments and safety cases rest on (Weij et al. 2024). The threat comes in at least two mechanically distinct forms. In prompted sandbagging the underperformance is induced by instructions in the context window and vanishes when the instructions do. In password-locked sandbagging the underperformance is trained into the weights: the model performs poorly by default and recovers its capability only when a trigger string is present in the prompt (Greenblatt et al. 2024), the same trained-in conditionality that lets deceptive behavior persist through safety training (Hubinger et al. 2024). A monitor that catches one form is not thereby a monitor of the other, and the trained-in form is the one an adversary or a misaligned training process would produce.

Linear probes are the standard cheap instrument for reading internal state of this kind. Intermediate activations carry linearly decodable structure (Alain and Bengio 2016), includ-

Preprint. Under review.

ing truthfulness and belief signals (Azaria and Mitchell 2023; Burns et al. 2022; Marks and Tegmark 2023), high-level behavioral dispositions (Zou et al. 2023), and strategic deception specifically, where probes reach high AUROC on held-out data from their training distribution (Goldowsky-Dill et al. 2025). This suggests an attractive monitoring recipe: elicit the bad state on demand, in this case by instructing the model to sandbag, train a probe on the resulting activations, and deploy it against the forms of the state that cannot be elicited on demand. The recipe rests on an assumption that is rarely tested, that the elicitable and the trained-in forms of the behavior share an encoding the probe can read.

This paper tests that assumption directly. In two registered 7-8B instruct families, Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, at six seeds each, we train linear probes only on prompted-sandbagged versus honest-best-effort activations from the base checkpoint and evaluate them on password-locked organisms under a neutral prompt with no sandbag instruction in context, in both transfer directions. Three registered gates make a transfer failure interpretable: a target-trained oracle probe must detect the locked state (so failure is not information absence), a behavior-neutral LoRA control must preserve retention (so failure is not generic checkpoint drift), and the cosine between the two mean-difference directions separates distinct encodings from a shared direction with a shifted threshold. A post-registration replication arm runs the same pipeline in Mistral-7B-Instruct-v0.3.

The probes are near ceiling in distribution in every run, and every gate holds, yet cross-condition transfer is weak and frequently sign-inverted, in nineteen of twenty-four registered family-seed-direction runs. The registered retention hypothesis is not confirmed as stated; the one claim that survives family-wise correction is a pooled asymmetry, locked-to-prompted retention exceeding prompted-to-locked retention. A masked-oracle control shows the oracle's near-perfect detection tracks the locked state rather than the password string, and direction patching in a validated final-layer regime shows the sandbagging directions carry modest, direction-specific causal leverage in three of four family-direction combinations. The contribution is not a new detector. It is a controlled measurement of which kind of sandbagging a prompted probe actually covers, and the answer, that in-distribution validation can sit at ceiling while cross-mechanism transfer inverts, bears directly on how probe-based monitoring should be validated before deployment.

Related work

Sandbagging and password-locked organisms. Strategic underperformance in language models was established for both prompted and fine-tuned mechanisms by van der Weij et al., who show open models can be made to selectively underperform on evaluations (Weij et al. 2024). Password-locked models were developed as a stress test for capability elicitation: Greenblatt et al. train models whose capability is gated on a trigger string and study what it takes to elicit the hidden capability (Greenblatt et al. 2024). The sleeper-agents line shows that trained-in conditional misbehavior of this shape persists through standard safety training, which is what makes the weight-level form the concerning one (Hubinger et al. 2024). Our organisms follow this construction: multi-task LoRA locking (Hu et al. 2021) over MMLU and ARC-Challenge (Hendrycks et al. 2020; Clark et al. 2018), with CommonsenseQA reserved as a held-out domain (Talmor et al. 2018).

Probe-based detection of hidden state. That intermediate activations are linearly decodable is an old observation (Alain and Bengio 2016), since extended to internal truthfulness signals (Azaria and Mitchell 2023), latent knowledge extracted without supervision (Burns et al. 2022), and high-level behavioral dispositions read and steered as directions (Zou et al. 2023). Marks and Tegmark map the linear geometry of true and false statements and validate it interventionally (Marks and Tegmark 2023). Closest to our target, Goldowsky-Dill et al. train linear probes for strategic deception and report strong in-distribution performance (Goldowsky-Dill et al. 2025). This literature largely establishes decodability within a distribution; it is the premise of the monitoring recipe our study stress-tests, not evidence for its cross-mechanism step.

Probe generalization under mechanism shift. Levinstein and Herrmann show that internal-state lie detectors can fail to generalize even to modest distribution shifts (Levinstein and Herrmann 2023), and the sleeper-agents work observes that detection tuned to one trigger

regime need not carry to another (Hubinger et al. 2024). What this strand lacks is a controlled test where the behavior is held fixed while only the eliciting mechanism changes, with positive controls establishing that the target state is decodable at all.

That controlled test is the delta of this paper. We train probes only on prompted sandbagging from the base checkpoint and evaluate them on password-locked sandbagging in the same family under a neutral prompt, in both directions, with the failure claim triple-gated: a target-trained oracle proves the locked state is decodable, a behavior-neutral LoRA control rules out generic checkpoint drift, and the mean-difference cosine separates distinct encodings from a shared, rethresholded direction. The result is not a new detector but a measurement of coverage, which kind of sandbagging a cheap prompted probe actually monitors, together with the observation that the transfer that does survive is frequently sign-inverted rather than merely attenuated.

Methods

Families and scope. The registered hypothesis is defined on two 7-8B instruct families, Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, each realized at six seeds through the same locking, behavior-neutral control, extraction, probing, and analysis stages, and nothing in this section alters that registered scope. The design of record additionally specifies a post-registration replication arm in a third family, Mistral-7B-Instruct-v0.3, at the same six-seed count and pipeline (design.json, scale.mistral_extension); the extension leaves the registered two-family hypothesis block unchanged and its ten registered tests are reported separately in Results as a replication arm rather than as part of the registered claim. Statements about the registered hypothesis therefore concern Qwen and Llama, while the realized study spans three families.

Analysis provenance. The design of record pre-specified the retention criterion in both transfer directions, the three gate claims (target-trained oracle AUROC of at least 0.80, behavior-neutral LoRA retention of at least 0.70, mean-difference cosine below 0.30), the pre-answer token position as primary with mean pooling over the last eight tokens and the last-instruction token as position ablations, per-seed layer selection on prompted-validation AUROC only, a plan of three seeds per family with the decision rule defined at that count (design.json, seeds), and cross-condition direction patching at the registered settings (300 items per direction) as the causal instrument. The realized study extends each family to six seeds; that extension is post-registration, is reported without re-registering the decision rule, and reverses the Llama prompted-to-locked verdict the three-seed mean had satisfied. The following analyses are likewise post-registration and should be read as exploratory or remedial rather than pre-specified: the sign-aware retention reanalysis, introduced after inverted transfer was observed; the re-extraction pass for the registered position ablations, remedial because the committed activations supported only the primary position; the patching magnitude sweep, the letter-logit positive control, and the matched cross-condition patching in the validated final-layer norm-scaled regime, adopted after the registered instrument failed its own positive control; and the password-masked oracle control. The Mistral-7B-Instruct-v0.3 replication arm was mandated by the wave-5 design review as a post-registration extension (design.json, scale.mistral_extension) and is reported separately from the registered claim throughout.

Results

Sign-aware retention. The pre-registered retention metric takes the absolute distance of the transfer AUROC from chance, R = |AUROC_transfer - 0.5| / (AUROC_in-dist - 0.5), and therefore cannot distinguish residual signal that points the right way from residual signal that points the wrong way. A sign-aware reanalysis recomputes signed retention, signed_R = (AUROC_transfer - 0.5) / (AUROC_in-dist - 0.5), which preserves the direction of the retained signal; negative values flag transfer with inverted sign (sign_aware_retention.csv).

Regenerated over all six seeds per family, the sign-aware analysis covers twenty-four family-seed-direction runs; the residual signal is inverted in nineteen and aligned in five, the aligned cases being both directions in Llama seeds 0 and 4 and the locked-to-prompted direction in Llama seed 1. For Qwen every run is inverted. Prompted-to-locked signed_R is -0.327,

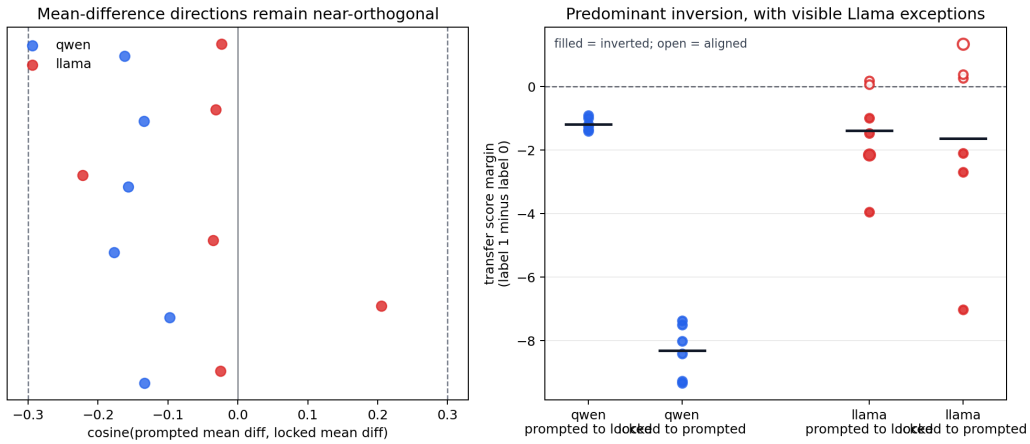

Figure 1: Subspace geometry of prompted vs password-locked sandbagging: near-zero mean-difference cosines alongside projection margins showing the sign-inverted transfer.

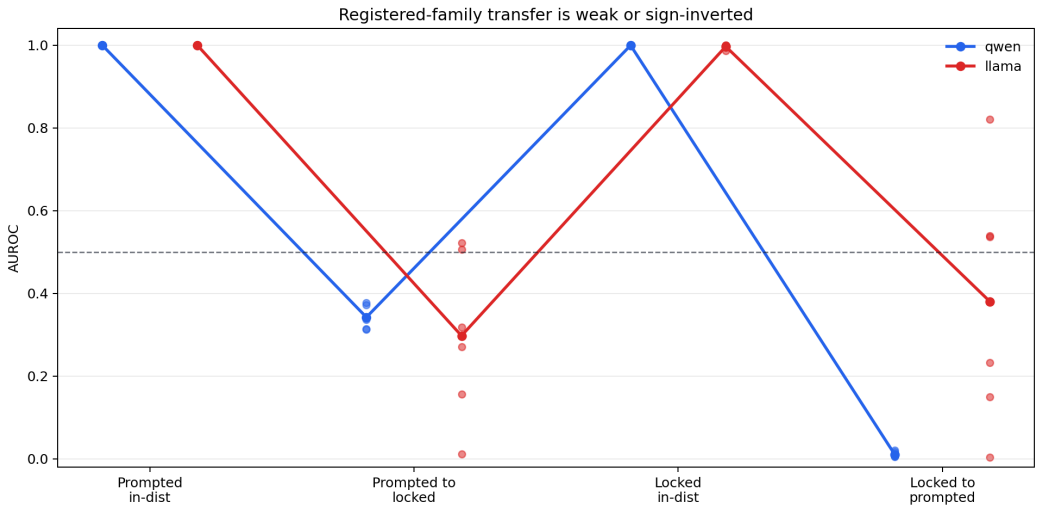

Figure 2: Transfer AUROCs by family and direction. In-distribution probes remain near-perfect, but prompted-to-locked transfer is below chance and locked-to-prompted is inverted for Qwen.

-0.245, and -0.373 across seeds 0 to 2, at transfer AUROCs of 0.336, 0.378, and 0.313 against in-distribution AUROCs of 0.998 or higher, and locked-to-prompted signed_R is -0.988, -0.962, and -0.988, at transfer AUROCs of 0.007, 0.021, and 0.006. For Llama, seed 0 retains aligned signal of low magnitude in both directions (signed_R 0.045 prompted-to-locked and 0.074 locked-to-prompted), seed 1 is inverted prompted-to-locked (-0.686) but aligned locked-to-prompted (0.642), and seed 2 is inverted in both directions (-0.364 and -0.535).

The two readings of the criterion therefore come apart. Under the pre-registered absolute criterion, absolute R below 0.40, the prompted-to-locked direction passes in six of six Qwen seeds and three of six Llama seeds. A sign-aware criterion that additionally requires the retained signal to be aligned, $0 <= \text{signed\_R} < 0.40$, is satisfied in only four of the twenty-four runs, both directions in Llama seeds 0 and 4, and fails everywhere else through inversion, magnitude, or both. In particular, the Qwen prompted-to-locked passes under the absolute criterion rest on inverted residual signal: what little discrimination the prompted probe retains on the locked organism points in the wrong direction in all six seeds.

Layer sweep. The probe layer was selected per family and seed by a sweep of prompted-validation AUROC over all layers, and the full curves show that this criterion provides almost no selection signal (full_layer_sweep_curves.csv). Across all 360 layer points (32 layers in Llama and 28 in Qwen, six seeds each), prompted-validation AUROC never falls below 0.9958, and 353 of 360 points lie at 0.999 or above, with the only values below 0.999 occurring at layers 0 or 1. Every per-seed curve peaks at 0.99998 or higher, so the selected layers (1, 25, 4, 6, 1, and 4 for Llama seeds 0 to 5, and 2, 2, 3, 2, 2, and 2 for Qwen) are chosen by margins at the fourth decimal place or beyond among layers that are all effectively at ceiling. The variation in selected layer across seeds, most visibly Llama seed 1 selecting layer 25 where its sibling seeds select early layers, therefore reflects near-ties among saturated values rather than meaningful localization of the sandbagging signal, and the in-distribution ceiling holds at every depth in both families.

Position ablation. The design registers the pre-answer token position as primary, with mean pooling over the last eight tokens and the last-instruction token serving as a position ablation. An audit of the committed activations found that this ablation cannot be evaluated from the run's artifacts (position_ablation.csv). The committed tensors under acts/real have shape (items, layers, hidden) with no position axis, matching only the pre-answer extraction path (manifest shapes (2018, 32, 4096) and (2337, 28, 3584), 21 tensors each); no mean-pooled or last-instruction-token tensor was committed for any family or seed, and probes for those positions cannot be trained without a re-extraction pass. The pre-answer probes are present, and the audit's AUROCs for them match the values reported above (for example, Qwen locked-to-prompted at 0.007, 0.021, and 0.006 across seeds). A dedicated re-extraction pass subsequently closed this gap; its results follow.

Position ablation results. The re-extraction produced mean-pool-last-8 and last-instruction-token activations for all six conditions in every family and seed, and probes were trained and evaluated at both positions with the same pipeline as the primary analysis (position_ablation_wave2.csv, 24 position rows; per-item scores in position_probe_scores_wave2.csv, 96672 rows). The headline transfer behavior is not an artifact of the pre-answer position. At the last-instruction token the inversion is the most extreme observed anywhere in the study: transfer AUROCs are 0.013 or below in both directions in every family and seed, giving signed retention between -0.97 and -1.01, while in-distribution, oracle, and neutral-LoRA AUROCs stay at 0.992 or above; the sign-aware criterion is satisfied in zero of the twenty-four direction-runs at this position. Mean pooling over the last eight tokens is the mildest of the three positions. There Llama sits near chance in most seeds (signed retention from -0.17 to +0.30, satisfying the sign-aware criterion in six of twelve direction-runs), while Qwen remains inverted, mildly so prompted-to-locked (signed retention -0.02 to -0.31) and strongly so locked-to-prompted in five of six seeds (-0.64 to -0.95, with seed 2 the exception at +0.06), satisfying the sign-aware criterion in one of twelve. The validation-selected layers are again early (1 to 7 for mean pooling, layer 0 in every last-instruction run), and in-distribution AUROC is at ceiling at every position, so the position ablation reproduces the study's central pattern of near-perfect in-distribution detection alongside transfer that is absent or sign-inverted at the positions and layers the pre-registered selection procedure picks.

Threshold geometry. A reconciliation analysis places the direction cosines and the probe score distributions for each transfer side by side (bias_threshold_geometry.csv). The mean-difference cosine between the prompted and locked sandbagging axes is small in every seed (-0.025, 0.205, and -0.036 for Llama seeds 0 to 2; -0.134, -0.097, and -0.177 for Qwen), yet the transfer scores are not scrambled. Under the locked probe applied to prompted data, honest items score well above sandbag items in every Qwen seed, with label-0 score means of 6.80, 10.76, and 11.09 against label-1 means of -1.61, 3.40, and 1.76, giving the margins of -8.41, -7.36, and -9.33 behind the inverted AUROCs reported above. The Llama margins are smaller and mixed in sign across seeds, for example 0.180 and 0.261 in the two directions of seed 0 but -2.14 and 1.33 in seed 1. The analysis notes that near-orthogonal mean-difference axes are compatible with strong, even inverted, transfer because the trained probe is an affine classifier whose direction and threshold can order the target labels consistently, or consistently reversed, along a direction distinct from the mean-difference axis.

Direction patching controls. Two controls calibrate the direction-patching instrument at the same settings as the main patching analysis, alpha of 1.0 and 300 items per cell (causal_patch_controls.csv, 7200 item rows summarized in causal_patch_controls_summary.csv). The self-transfer positive control injects each condition's own mean-difference direction into that same condition, prompted mean-difference on prompted honest items and locked mean-difference on locked with-password items, where an instrument with real behavioral leverage should visibly shift answers. The random-direction null injects a random direction at the same alpha. Both come back near zero. In Qwen the positive-control flip rate never exceeds 0.003 in any of the six cells, indistinguishable from its null cells of 0.000 to 0.003. In Llama the positive control ranges from 0.000 to 0.063 (the largest cell being prompted honest at seed 1, where patched accuracy drops from 1.000 to 0.937) against null cells of 0.003 to 0.027, so it exceeds the null in some seeds but never moves more than 7% of answers. Across all 24 cells, patched accuracy differs from clean accuracy by at most 0.063. The positive control therefore fails to demonstrate instrument sensitivity: at this alpha, injecting even a condition's own sandbagging direction leaves the overwhelming majority of answers unchanged, and small flip rates from mean-difference direction patching cannot be read as evidence for or against a causal role of these directions.

Patching magnitude sweep. A revised sweep re-runs direction patching at magnitudes alpha of 1, 2, 4, and 8 with 180 items per cell, covering two same-direction positive controls, two cross-direction patches, and two random-direction nulls per family and seed (wave2_causal_patch_sweep.csv, 51840 patched items summarized in 288 cells in wave2_causal_patch_sweep_summary.csv). For Qwen the instrument stays inert at every magnitude: across all 144 Qwen cells the flip rate never exceeds 0.022, so raising alpha does not rescue the failed positive control and no causal conclusion about the Qwen directions is available from patching. For Llama the sweep separates one seed from the rest. In five of six seeds the directional patches never flip more than 0.128 of answers at any alpha, while the random nulls grow nonspecifically with magnitude and flip 0.133 to 0.439 of answers at alpha 8, meaning that in those seeds matched random directions disrupt behavior at least as much as the sandbagging directions. Seed 1 is the exception, and there the effect is direction-specific: the locked mean-difference direction flips 0.683, 0.772, and 0.606 of answers on the locked organism at alphas 2, 4, and 8, and 0.300 to 0.667 on prompted-sandbag items, with the prompted direction reaching 0.844 on prompted items at alpha 8, while the same seed's random locked null flips 0.000 at every alpha and its random prompted null at most 0.061. Llama seed 1 is the same seed that is atypical throughout this section, selecting layer 25 where its siblings select early layers and showing aligned rather than inverted locked-to-prompted transfer; it is also the only run in either family in which the mean-difference directions show behaviorally potent, direction-specific causal leverage.

Cross-patch versus null statistics. The magnitude sweep's comparison of cross-direction patches against their random-direction nulls is also tested formally, with paired sign-permutation tests over the 48 matched family, seed, condition, layer, and alpha cells per direction and bootstrap confidence intervals over the paired flip-rate differences (causal_patch_vs_null_stats.csv, computed from wave2_causal_patch_sweep_summary.csv). Neither direction shows a statistically supported direction-specific effect. Injecting the locked direction into prompted items exceeds its matched null by a mean flip-rate difference of 0.037 (CI -0.006 to 0.089, raw $p = 0.07$, Holm-corrected 0.94, BH-corrected 0.126), a positive trend consistent with the Llama seed 1 concentration described above but with a confidence interval that includes zero. Injecting the prompted direction into locked items does not exceed its null at all, with a mean difference of -0.017 (CI -0.036 to +0.003, raw $p = 0.956$, corrected 1.0 under both procedures), meaning the matched random directions flip answers at least as often as the sandbagging direction. The formal test therefore confirms the qualitative reading of the sweep: across families and magnitudes, mean-difference direction patching produces no statistically supported direction-specific behavioral effect in either cross direction.

Letter-logit positive control. The instrument calibration above left open whether direction patching at any accessible setting could detect a real effect, so a followup control injects a direction with known behavioral leverage, the direction that moves the model's answer-letter logit, as a unit direction scaled per item by the activation norm

at the injection layer (causal_patch_positive_control.csv, 34560 patched items summarized in causal_patch_positive_control_summary.csv). The known direction is injected on prompted honest items at the registered best layers and the final layer, at norm multiples 0.25 to 2.0 with 180 items per cell, alongside a norm-matched random-direction null. At the final layer the known direction flips every answer in every cell, a flip rate of 1.000 in all 24 cells per family (layer 31 in Llama, layer 27 in Qwen), while the norm-matched null reaches at most 0.617 in Llama and 0.067 in Qwen. At the non-final, registered-best layers the same known-good direction averages a flip rate of 0.093 pooled across families (0.180 in Llama, 0.006 in Qwen), in the range the mean-difference patches produced above. The control therefore validates the patching instrument in one specific regime, final-layer injection at activation-norm scale, and shows that a direction with known behavioral leverage moves few answers at the earlier layers the wave-2 sweep used. The wave-2 cross-condition readings above are unchanged by this; the validated final-layer regime is new territory, and the matched cross-condition patch that follows was run in it.

Matched cross-condition patching. The cross-condition patches were then re-run inside the validated regime (run r-735fd3dd4c63): each family's prompted and locked mean-difference directions are injected as unit directions scaled per item by the activation norm at the injection layer, on 180 items per cell at the registered best layer and the final layer, at norm multiples 0.25 to 2.0 over six seeds per family, against random-direction nulls matched cell for cell, 69120 patched items in all (causal_patch_matched.csv, causal_patch_matched_stats.csv). Three of the four family-direction combinations now exceed their matched nulls. Injecting the prompted direction into locked-with-password items raises the flip rate over the null by a mean of +0.045 in Llama (flip rate 0.164 against null 0.119; bootstrap CI 0.011 to 0.084, sign-permutation $p = 0.0086$ over the 48 matched cells) and +0.077 in Qwen (0.085 against 0.008; CI 0.039 to 0.121, $p < 1e\text{-}5$). Injecting the locked direction into prompted-sandbag items exceeds the null by +0.070 in Llama (0.334 against 0.265; CI 0.030 to 0.112, $p = 0.00066$) but not in Qwen, where the difference is +0.005 (0.079 against 0.074; CI -0.010 to 0.021, $p = 0.27$) and the null hypothesis stands. Pooled over families the differences are +0.061 (CI 0.035 to 0.089, $p = 1e\text{-}5$) prompted-to-locked and +0.037 (CI 0.015 to 0.061, $p = 0.00073$) locked-to-prompted. The effects are direction-specific but modest, mean differences of 0.05 to 0.08 where the positive control flips everything, and they are asymmetric in Qwen, whose locked direction moves prompted behavior no better than a random direction. Under an instrument regime with demonstrated sensitivity, the sandbagging mean-difference directions therefore do carry some causal leverage across conditions in three of four cases, in contrast to the earlier sweep where no setting could have detected it.

Masked-oracle control. The oracle probe behind the gate claims is trained on locked-organism activations from prompts that contain the password literal, leaving open the possibility that it reads the password string rather than the locked state. A control re-extracts locked-organism activations with the password replaced by a neutral filler (run r-012a88dd38f9; password_masked_oracle_summary.csv, 12 family-seed rows). Because a fixed literal such as "[MASKED]" changes the token count (-2 in Llama, -4 in Qwen), the filler is length-matched per family tokenizer, with the alignment measured rather than assumed: zero token-count and zero pre-answer-position deltas in both families on the sampled items (password_masked_oracle_token_alignment.json). On masked activations the oracle still separates the locked organism from honest at AUROC 0.9916 in Llama (bootstrap CI 0.983 to 1.000) and 0.9806 in Qwen (CI 0.975 to 0.986), above chance at the attainable seed-level floor ($p = 0.016$ in both families), with a small decrease against the unmasked held-out oracle (-0.008, $p = 0.0625$ in Llama; -0.019, $p = 0.031$ in Qwen). The oracle signal therefore tracks the locked state rather than the password token sequence. Behaviorally the masked prompts sit between the two locked conditions: the locked organism answers at accuracy 0.487 (Llama) and 0.559 (Qwen) under masking, against 0.074 and 0.005 with no password and 0.963 and 0.969 with the password present, so masking partially but not fully withholds the unlock.

Additional seeds. Seeds 3 to 5 were run through the full pipeline for both families, bringing each family to six seeds (metrics_raw.csv and analysis_summary.txt, seeds_observed 0 to 5). Qwen replicates its original pattern in every new seed: prompted-to-locked AUROCs of 0.314, 0.372, and 0.340 (R of 0.371, 0.255, and 0.319, all below 0.40 and all below chance),

locked-to-prompted AUROCs of 0.005, 0.017, and 0.006 (R of 0.990, 0.966, and 0.988), and mean-difference cosines of -0.157, -0.134, and -0.162. Over six seeds the Qwen family means are R = 0.315 prompted-to-locked and R = 0.980 locked-to-prompted at AUROC 0.010, with cosine -0.144; all six prompted-to-locked runs fall below the 0.40 criterion and all six transfer with inverted sign. Llama, by contrast, grows more unstable as seeds accumulate. Seed 4 behaves like seed 0, with near-chance transfer in both directions (AUROCs 0.506 and 0.539, R of 0.013 and 0.079). Seed 3 behaves like Qwen and more extremely so, with near-perfectly inverted transfer in both directions (AUROCs 0.011 and 0.004, R of 0.977 and 0.992) and a held-out-domain transfer AUROC of 0.0015. Seed 5 sits between, at AUROCs of 0.271 and 0.149 (R of 0.458 and 0.701). The six-seed Llama tallies are three of six seeds below 0.40 in the prompted-to-locked direction (seeds 0, 2, and 4) and two of six in the locked-to-prompted direction (seeds 0 and 4), and the Llama family mean prompted-to-locked R rises to 0.424, above the criterion value its three-seed mean satisfied. The pre-registered decision rule was defined for three seeds per family, so we report the six-seed tallies without re-registering it. Pooled over all twelve runs, prompted-to-locked R is 0.369 (95% bootstrap CI 0.240 to 0.527) and locked-to-prompted R is 0.742 (CI 0.538 to 0.917) at AUROC 0.195 (CI 0.062 to 0.359), and the held-out domain pools to AUROC 0.320 (CI 0.244 to 0.390). The gates hold in every one of the twelve runs: oracle AUROC is at least 0.9866, neutral-LoRA retention at least 0.9968, and the absolute mean-difference cosine at most 0.222 (Llama seed 3), all on the satisfying side of their criteria.

Wave 1 aggregate and verification. All diagnostic analyses above were regenerated over the full six-seed grid (sign_aware_retention.csv, random_direction_floor_bounded.csv, full_layer_sweep_curves.csv, position_ablation.csv, bias_threshold_geometry.csv, manipulation_accounting.csv), and the finalize step confirms the aggregate covers 12 metric rows, both families, seeds 0 to 5, and 13 statistical claims (wave1_finalize_report.json, stats.json). At six seeds per family the smallest attainable seed-level sign-permutation p is 0.016. The Qwen prompted-to-locked criterion holds at that floor (mean R = 0.315, p = 0.016) while the Llama mean of 0.424 does not (p = 0.547), and the locked-to-prompted direction fails in both families (Qwen mean R = 0.980, p = 1.0; Llama 0.504, p = 0.75). The transfer asymmetry, locked-to-prompted retention minus prompted-to-locked retention, is 0.373 pooled across all twelve runs (p = 0.001), 0.665 in Qwen (p = 0.016), and 0.080 in Llama (p = 0.0625). Every gate claim holds at the attainable floor in both families: oracle AUROC 0.999 (Qwen) and 0.997 (Llama), neutral-LoRA retention 1.000 and 0.999, and mean absolute cosine 0.144 and 0.090, each at p = 0.016. Finally, an independent recompute of the printed numbers in analysis_summary.txt and the statistics in stats.json from the committed raw artifacts reproduced them with zero mismatches (recompute_retest: summary_mismatch_count 0, stats_mismatch_count 0), so the numbers in this section are verified end-to-end against the raw per-item scores.

Multiple-comparisons correction. The thirteen registered seed-level tests are corrected jointly with Holm-Bonferroni and Benjamini-Hochberg FDR procedures, with the correction family being all thirteen tests (stats.json). The correction interacts sharply with the permutation floor: at six seeds the smallest attainable raw p is 0.016, so the eight tests sitting at that floor (the Qwen prompted-to-locked criterion, the Qwen transfer asymmetry, and the six gate claims) share a Holm-corrected p of 0.188 and a BH-corrected p of 0.023. Under BH at the 0.05 level, nine of the thirteen tests survive, those eight plus the pooled transfer asymmetry (raw p 0.001, corrected p 0.013 under both procedures). Under the stricter family-wise Holm correction, the pooled transfer asymmetry across all twelve family-seed runs is the only claim that survives at 0.05. The four remaining tests are non-significant under either procedure: Llama prompted-to-locked (BH-corrected 0.646), Llama locked-to-prompted (0.813), the Llama transfer asymmetry (0.081), and Qwen locked-to-prompted (1.0, the direction in which the criterion fails outright). Statistical certification of the individual per-family claims is therefore limited by seed count rather than by effect size, and the single finding robust to family-wise correction is that locked-to-prompted retention exceeds prompted-to-locked retention across families.

Third-family replication. The design of record specifies a post-registration replication arm in Mistral-7B-Instruct-v0.3 at the same six-seed count and pipeline (design.json, scale.mistral_extension), reported here separately from the registered two-family claim as

Methods states. Its registered statistics live in stats.json with printed values in analysis_summary.txt (mistral_ lines), and an independent recompute of all 60 statistics entries in the registry from the committed raw artifacts reproduced them with zero mismatches (wave6b_finalize_report.json). The gates replicate. The oracle probe detects the locked state at AUROC 0.9996 (criterion at least 0.80, p = 0.016), the behavior-neutral LoRA control retains R = 0.9995 (criterion at least 0.70, p = 0.016), the absolute mean-difference cosine is 0.125 (criterion below 0.30, p = 0.016), and the prompted probe validates in-distribution at AUROC 0.99999. The transfer pattern is weaker and less directional than in either registered family. Mean absolute retention is R = 0.411 prompted-to-locked (p = 0.547 against the 0.40 criterion) and R = 0.397 locked-to-prompted (p = 0.484), at transfer AUROCs of 0.551 and 0.547 and a held-out-domain AUROC of 0.586; the transfer asymmetry that separated the registered families is absent at -0.014 (p = 0.594). The absolute retention reflects sign instability rather than aligned signal: signed retention is 0.102 prompted-to-locked with three of six seeds inverted and 0.093 locked-to-prompted with two of six inverted, and the transfer signal exceeds the bounded random-direction floor in the locked-to-prompted direction (mean excess 0.110, p = 0.047) but is not certified in the other (0.119, p = 0.078). Neither retention criterion is certified in Mistral. The family resembles Llama's seed-heterogeneous behavior more than Qwen's uniform inversion, so the replication arm reproduces the study's core contrast, near-ceiling in-distribution, oracle, and control probes alongside weak, sign-unstable cross-condition transfer, without meeting the registered retention criterion that Qwen met.

Discussion

The central result is not that prompted and password-locked sandbagging share no structure; it is narrower and more useful: a linear probe trained in one elicitation regime was not a reliable monitor of the other, even while in-distribution probes, oracle probes, and control probes all performed near ceiling. In every family-seed run across three families a linear probe trained and validated on prompted sandbagging separates its own distribution essentially perfectly, and in all twelve registered runs the target-trained oracle detects the password-locked state at AUROC 0.9866 or higher, the behavior-neutral LoRA control retains at least 0.9968 of its discrimination, and the absolute mean-difference cosine between the prompted and locked directions stays at or below 0.222; the Mistral arm replicates all three gates at the family level. The preconditions of the registered hypothesis are therefore met everywhere: both forms of underperformance are linearly detectable, the checkpoint drift explanation is ruled out, and the two axes are close to orthogonal. Against that backdrop, cross-condition transfer fails, but it fails in a way the pre-registered retention metric was not designed to describe. Because R takes the absolute distance from chance, it cannot distinguish residual signal that points the right way from signal that points the wrong way, and the sign-aware reanalysis shows that nineteen of the twenty-four registered-family direction-runs transfer with inverted sign. Stated plainly, the registered hypothesis is not confirmed: it required retention below 0.40 in both directions in both families, Qwen fails it outright in the locked-to-prompted direction (R = 0.980), the Llama prompted-to-locked mean sits above the criterion at six seeds (R = 0.424), and the per-family passes it does contain rest on residual signal that points the wrong way in every Qwen seed. What the sign-aware reading supports instead is the phenomenon the hypothesis was aimed at, in a form stronger than the metric anticipated: the two forms of underperformance are encoded distinctly enough that a probe carried across conditions is not merely weakened but often reversed. At the level of statistical certification, the picture is limited by seed count rather than effect size: at six seeds the smallest attainable permutation p is 0.016, and after family-wise Holm correction the only surviving claim is the pooled transfer asymmetry, locked-to-prompted retention exceeding prompted-to-locked retention by 0.373 across the twelve registered runs (p = 0.001). The per-family criterion results survive the false-discovery-rate correction but not the family-wise one, and we report them at that strength and no more.

The Qwen asymmetry deserves plain statement because it is the strongest single pattern in the study and also the least intuitive. A probe trained on Qwen's password-locked organism, applied to prompted sandbagging, does not degrade toward chance; it orders the classes almost perfectly backwards, at transfer AUROCs between 0.005 and 0.021 across all six

seeds. The threshold geometry analysis reconciles this with the near-zero mean-difference cosines: an affine probe's direction and threshold can order target labels consistently reversed along a direction distinct from the mean-difference axis, so orthogonal axes and strong inverted transfer are compatible. The position ablation shows the same structure is not an artifact of the pre-answer extraction point, since the inversion is even more extreme at the last-instruction token, where signed retention sits near -1.0 in both registered families in every seed. The Mistral replication arm, by contrast, shows nothing this directional, and we report its nulls as nulls. Mistral replicates every gate (oracle AUROC 0.9996, neutral-LoRA retention 0.9995, cosine 0.125) but certifies neither retention criterion, its transfer AUROCs sit near chance at 0.551 and 0.547 with sign flipping across seeds, and the transfer asymmetry that separates the registered families is absent there (-0.014, $p = 0.594$). The honest summary is that the core contrast, near-ceiling in-distribution and oracle detection alongside weak or inverted cross-condition transfer, reproduces in a third family, while the specific quantitative criteria registered for Qwen and Llama do not.

The causal evidence is deliberately scoped, and its provenance should be stated as plainly as its statistics. The initial direction-patching analyses were uninterpretable because the instrument itself failed its positive control: at the registered layers and magnitudes, injecting even a condition's own sandbagging direction moved almost no answers, and matched random directions disrupted behavior as much as the sandbagging directions did. The letter-logit control located a regime in which the instrument demonstrably works, final-layer injection at activation-norm scale, where a direction with known behavioral leverage flips every answer in every cell. That regime was identified only after the registered instrument failed, so the causal analysis is exploratory in provenance, and only inside the validated regime do we make causal claims. There, the matched cross-condition patches exceed their random-direction nulls in three of four family-direction combinations, in Llama in both directions (+0.045, $p = 0.0086$, and +0.070, $p = 0.00066$) and in Qwen prompted-to-locked (+0.077, p below 1e-5), while Qwen's locked direction moves prompted behavior no better than a random direction (+0.005, $p = 0.27$), mirroring the extreme inversion of the corresponding probe transfer. The effect sizes are modest, mean flip-rate differences of 0.05 to 0.08 where the positive control flips everything. This is evidence that the linear directions are not merely correlational readouts; it is not evidence that they are the model's mechanism for switching between honest and sandbagged behavior. Probe portability and causal manipulability are related questions, and these results show they are not interchangeable. We do not know whether these directions have leverage at earlier layers, only that this instrument cannot detect it there.

The masked-oracle control closes the most obvious deflationary reading of the gate claims. The oracle is trained on prompts that contain the password literal, so it could in principle be reading the token sequence rather than the locked state. With the password replaced by a per-tokenizer length-matched filler, verified to produce zero token-count and zero pre-answer-position deltas, the oracle still separates the locked organism from honest at AUROC 0.9916 in Llama and 0.9806 in Qwen, with a small decrease against the unmasked held-out oracle that reaches significance in Qwen (-0.019, $p = 0.031$) but not in Llama (-0.008, $p = 0.0625$). The locked-state signal, in other words, is nearly all state and very little string. Behaviorally the masked prompts sit between the two locked conditions, so masking partially withholds the unlock, which makes the near-intact oracle performance the more notable.

Several limitations bound these conclusions. The registered claim covers exactly two model families at the 7-8B scale, and the Mistral arm, though run at the same seed count through the same pipeline, is a post-registration replication and is reported as such. The pre-registered decision rule was defined for three seeds per family, and its verdict is not stable under extension: the Llama prompted-to-locked mean satisfied the criterion at three seeds and sits above it at six, so we report the six-seed tallies without re-registering the rule. Probe layers are selected on prompted-validation AUROC that is saturated everywhere (353 of 360 layer points at 0.999 or above), so the registered selection procedure carries almost no information about where the signal lives. All probing is linear, sandbagging is elicited in a single task domain with one held-out domain for transfer, and the password-locked organisms are produced by one locking procedure, so the results speak to this construction of hidden underperformance rather than to sandbagging in general. The causal analysis

is confined to the one patching regime the positive control validated. And Llama's seed heterogeneity, including one seed with direction-specific behavioral leverage that no sibling seed shows, is a reminder that per-seed idiosyncrasy at this scale can be large relative to the effects under study.

For sandbagging detection the practical implication is uncomfortable in a specific way. The failure mode we observe is not a probe that degrades to chance, which a monitoring pipeline would notice as low confidence, but a probe that remains confidently wrong, ordering the classes backwards while its in-distribution validation sits at ceiling. A monitor trained on instructed or prompted underperformance and deployed against a model whose underperformance is trained in could therefore be worse than uninformative, and nothing in its training-time validation would warn of it. The oracle results cut the other way and are the more hopeful finding: when even a modest amount of supervision from the target distribution is available, the locked state is linearly detectable almost perfectly, and the masking control says that detection does not lean on superficial prompt cues. The near-orthogonality of the two sandbagging axes suggests monitoring approaches that assume a single generalizing deception direction are, at least at this scale and for this pair of elicitation methods, resting on an assumption our data contradict. Future work should test whether the asymmetry and the inversion survive at larger scales and across other locking procedures, extend the task domains beyond the single registered one, re-register decision rules at seed counts that allow per-family certification, probe with richer function classes than affine ones, and ask directly what the inverted signal is tracking, since a consistently reversed ordering is itself structure, and structure that a monitor could in principle learn to use rather than be misled by.

References

Alain, Guillaume, and Yoshua Bengio. 2016. "Understanding Intermediate Layers Using Linear Classifier Probes." *arXiv Preprint arXiv:1610.01644*.

Azaria, Amos, and Tom Mitchell. 2023. "The Internal State of an LLM Knows When It's Lying." *arXiv Preprint arXiv:2304.13734*.

Burns, Collin, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. "Discovering Latent Knowledge in Language Models Without Supervision." *arXiv Preprint arXiv:2212.03827*.

Clark, Peter, Isaac Cowhey, Oren Etzioni, et al. 2018. "Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge." *arXiv Preprint arXiv:1803.05457*.

Goldowsky-Dill, Nicholas, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. "Detecting Strategic Deception Using Linear Probes." *arXiv Preprint arXiv:2502.03407*.

Greenblatt, Ryan, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. 2024. "Stress-Testing Capability Elicitation with Password-Locked Models." *arXiv Preprint arXiv:2405.19550*.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2020. "Measuring Massive Multitask Language Understanding." *arXiv Preprint arXiv:2009.03300*.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv Preprint arXiv:2106.09685*.

Hubinger, Evan, Carson Denison, Jesse Mu, et al. 2024. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." *arXiv Preprint arXiv:2401.05566*.

Levinstein, B. A., and Daniel A. Herrmann. 2023. "Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks." *arXiv Preprint arXiv:2307.00175*.

Marks, Samuel, and Max Tegmark. 2023. "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets." *arXiv Preprint arXiv:2310.06824*.

Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge." *arXiv Preprint arXiv:1811.00937*.

Weij, Teun van der, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. 2024. "AI Sandbagging: Language Models Can Strategically Underperform on Evaluations." *arXiv Preprint arXiv:2406.07358*.

Zou, Andy, Long Phan, Sarah Chen, et al. 2023. "Representation Engineering: A Top-down Approach to AI Transparency." *arXiv Preprint arXiv:2310.01405*.