The Boundary of a Behavioural Sandbagging Signature: Induced Token Concentration and Its Mechanism

JON-PAUL CACIOLI, Independent Researcher, Australia

A natural hope for detecting sandbagging, the deliberate underperformance of a language model on an evaluation, is that it leaves a behavioural trace, and prior work on small Llama-3 models reported one: under an explicit instruction to underperform, the answer distribution concentrates on a single option. We test whether this candidate signature generalises. Across ten models spanning five families and 7 to 72 billion parameters on 2,000 MMLU-Pro items, it appears in one of eight assessable models; several models comply with the instruction, dropping 25 to 40 accuracy points, without it, so its absence is not evidence of honest performance and a detector keyed to it would miss the strongest sandbaggers in the slate. It does not appear in the strongest non-Llama candidate, which leans toward serial position rather than label; it is Llama-lineage-specific here. Having bounded the phenomenon, we explain the one positive case. In Llama-3.1-8B the concentration follows the label token through a per-item position shuffle, so it is token-type not position-type, and it is at chance frequency in honest responding, so it is induced by the instruction rather than surfaced. An activation probe trained on honest trials decodes the correct answer from held-out sandbagging trials, including with the answer at a random position: the concentration is an output-side override of an intact content representation, well described by a Bernoulli override (rate 0.43) that reproduces the full accuracy drop. The contribution is a boundary result for a candidate detector, with a mechanistic account of its single positive case.

Additional Key Words and Phrases: sandbagging, strategic underperformance, selection bias, activation probing, evaluation validity

Preprint, July 2026. Pre-registered at https://osf.io/ykxp5. Not peer reviewed.

1 Introduction

Evaluations are trustworthy only if a model is not deliberately underperforming on them. Sandbagging, strategic underperformance, threatens this directly: a model that can score lower than its true capability can evade a capability threshold. Detecting sandbagging is therefore a prerequisite for trusting evaluation results.

One detection strategy looks for behavioural traces that distinguish genuine from feigned inability. A model that is truly uncertain spreads its answers across options in proportion to its partial knowledge. A model that is feigning may instead fall into a simpler pattern. Prior work in this programme found such a pattern in small Llama-3 models [7, 8]: under an explicit instruction to underperform, the distribution of emitted answer labels concentrates toward one option, regardless of which option is correct. The entropy of the answer distribution drops and one label absorbs a large share of responses. The prior programme called this positional collapse; we show below that the effect is not positional, and we therefore call it an induced concentration.

A candidate detector is only as good as the range over which it holds, so the first question is not how the trace works but whether it appears at all. This paper is organised around that question. We test the reported signature across a slate of ten models and find that it mostly does not generalise: it appears in one assessable model, and several models that comply with the instruction just as strongly, or more strongly, leave no trace at all. That

Author's Contact Information: Jon-Paul Cacioli, synthium@hotmail.com, Independent Researcher, Melbourne, Australia.

negative-generalisation result is the paper's primary contribution, because it speaks directly to whether the signature can be operationalised as evidence of honest performance (it cannot). Having established the boundary, we then explain the one case that falls inside it, characterising the mechanism of the concentration in the single positive model at both the behavioural and representational levels. The narrowness is not a caveat appended to a mechanism result; it is the result, and the mechanism is what the one exception turns out to be.

We use one term throughout for the phenomenon. We call it a concentration: under the sandbagging instruction the answer distribution concentrates on a single label. When we refer to the concentration in its role as a candidate detector, the thing one would look for to flag sandbagging, we call it the signature. We reserve "positional collapse" for the prior programme's name for it and avoid "fingerprint" and "attractor", the latter because it implies an internal dynamical mechanism the present behavioural evidence does not establish.

Concretely, across the slate the signature appears in one model, and several models comply with the instruction without any concentration; where it does appear, in Llama-3.1-8B, it is token-type not position-type, vocabulary-specific, and induced by the instruction rather than surfaced from a resting bias. The narrowness also makes the mechanism result well-defined: there is exactly one place to characterise it cleanly, and we do.

It helps to separate three axes that are easily conflated. Compliance is whether a model lowers its accuracy when instructed to underperform. Concentration is whether its answer distribution piles onto one label. Registration is whether a model gives any sign of having apprehended an incentive to underperform. These are distinct. A model can comply without concentrating, and our implicit-incentive condition tests registration rather than compliance or concentration. This tripartition is drawn from the validity-screening tradition discussed in Section 2.3, where performance validity, symptom validity, and the credibility of stated motivation are assessed as separate dimensions; it is not a post-hoc relabelling but the organising logic the programme imported from that field.

Three scope conditions bound every claim and we state them up front rather than in closing. The mechanism is characterised in one model. The phenomenon was first observed in the Llama lineage and the one positive here is also Llama, so a lineage-specific confound cannot be excluded. All conditions are single-turn multiple-choice with a deterministic decode; we do not test multi-turn, agentic, or sampled settings. These bounds are what the results are about as much as they are caveats to them.

This work sits at the intersection of three literatures: sandbagging and its detection, position and selection bias in multiple-choice evaluation, and validity screening in clinical and forensic assessment, from which our framing is drawn. Section 2 positions the contribution against each. One recent result partly anticipates ours and we engage it directly there.

2 Related Work

2.1 Sandbagging and Its Detection

Sandbagging, the strategic underperformance of a model on an evaluation, was characterised empirically by van der Weij et al. [28], who showed that models can be prompted or fine-tuned to underperform selectively and introduced password-locking as a model-organism technique. Subsequent work studied the generalisation of password-locked models [15] and situated sandbagging within broader treatments of scheming and deceptive alignment [18]. Frontier developers now report sabotage and capability-suppression evaluations [4], and

linear probes can separate evaluation from deployment prompts in a 70B open-weight model, with current safety evaluations classified as evaluations, suggesting they already read as artificial to models [20].

Detection approaches divide into black-box, white-box, and training-based methods. Black-box methods compare behaviour across framings or perturbations: noise injection into weights, which tends to raise a sandbagging model's performance [26], and answer-reasoning consistency checks. White-box methods probe internal representations, with linear deception probes reporting high AUROC on sandbagging scenarios [14], benchmarked by Parrack et al. [21]; recent work shows that linear probes can ride textual evidence and construction leakage in sandbagging model organisms rather than the target state [5], a concern the item-disjoint transfer protocol of Section 4.7 addresses directly. Training-based elicitation fine-tunes on a few demonstrations and can recover password-locked capabilities [15].

The concentration signature we study is a black-box, distributional method: it asks whether the answer-label distribution concentrates under induced underperformance. Our results bear on the limits of this family. Recent auditing-game work shows that black-box detection of sandbagging is not adversarially robust: a model can sandbag without signs in its chain of thought or outputs and can emulate a benign answer distribution to evade detection [25]. Separately, internal-probe work reports that evaluation-relevant features can be read from, and steered in, a sandbagging Llama-3.3-70B-Instruct, partially recovering suppressed capability [20]. Our contribution is complementary and prior in the natural order of questions: before asking whether a signature can be evaded, one should ask whether it appears at all, in which models, and by what behavioural mechanism. We find the signature in one model of ten, and absent in models that comply more strongly, including the same Llama-3.3-70B-Instruct that the internal-probe work studies. The dissociation we document, compliance without a distributional trace, is the behavioural counterpart of the evasion result: where the auditing-game work shows a signature can be hidden, we show it is mostly not present to begin with. A separate strand finds that sandbagging can be executed covertly against a chain-of-thought monitor, including by a Qwen2.5-72B model in the family of several of ours [17]; our reasoning-model observation is adjacent but does not address monitor-awareness directly.

2.2 Position, Selection, and Token Bias

The signature we test is a special case of a documented phenomenon. Models answering multiple-choice questions are not invariant to option order: they over-select options at particular positions [22] and over-favour particular option identifiers such as "A" [32]. The literature distinguishes two easily-conflated mechanisms. Position bias is a preference for a serial slot regardless of label; token bias, or selection bias, is a preference for a label token regardless of position [2, 32]. Mitigations include option-permutation averaging, input reformatting, and inference-time prior estimation such as PriDe [32].

Our label-remapping experiment is built on this distinction, and our result is stated in its terms: the concentration induced by the sandbagging instruction is token-bias-type, not position-bias-type. Two things make this more than a restatement of known selection bias. First, the concentration is induced by the instruction rather than present at rest: the preferred label is at chance frequency in honest responses and honest errors, where ordinary selection bias would already appear. Standard selection bias is a property of the model; the concentration we measure is a property of the model under a sandbagging instruction. Second, the preferred label is vocabulary-specific, differing across letters, numerals, and Greek, which a single fixed token prior does not predict. Recent work also reports that position bias in

reasoning models is trajectory-length-dependent, longer reasoning associated with more bias [29]; our observation that reasoning models shorten trajectories under sandbagging connects to this, since shortening would predict reduced bias, consistent with the flat distributions we observe.

2.3 Validity Screening in Clinical and Forensic Assessment

The framing of this programme is borrowed from clinical psychometrics, where the credibility of a response pattern is assessed before its content is interpreted. The MMPI-3 screens response validity (the L, F, and K scales) prior to interpreting clinical scales, and the PAI enforces a similar two-stage architecture [3, 19]. The logic is domain-independent: if a response pattern is dominated by a response set that overrides item content, downstream interpretation is fitting noise. We import this validity-first stance, treating answer-distribution entropy as a response-style indicator (analogous to the inconsistency and infrequency scales, which flag noncredible responding independent of content) and a difficulty-accuracy relationship as a content-engagement indicator (analogous to whether items discriminate as they should). These are not a single axis: in the MMPI tradition a profile can show an elevated infrequency scale without elevated response-inconsistency, and vice versa, and we find the parallel dissociation here, where a model can concentrate its responses while still tracking item difficulty.

We are deliberate about which feigning paradigm the conditions correspond to, because the clinical literature distinguishes them carefully and the mapping is easy to invert. The explicit sandbagging instruction is a simulation or coached-feigning design: the respondent is directly instructed to underperform, as in the instructed-malingering analogue studies used to validate symptom-validity measures, not malingering proper. Malingering in the clinical sense is feigning driven by an external incentive; that is closer to the implicit-incentive condition, which describes a consequence without instructing the behaviour. Factitious presentation, feigning from internal factors with no external incentive, is not modelled by any condition here. We therefore do not claim to capture malingering or factitious presentation as such; we use the instructed-feigning (simulation) paradigm as the well-validated base case and treat the incentive condition as a first step toward the externally-incentivised analogue. The performance-validity and symptom-validity test traditions [24, 27] supply the underlying principle, that noncredible responding can be detected from embedded indicators, rather than a literal taxonomy we map onto one-to-one; they are two channels for detecting noncredible presentation, not orthogonal dimensions, and we use them analogically. To our knowledge the transfer of this validity-screening logic to model evaluation has not previously been connected to the sandbagging-detection literature outside this programme [6–8], and we take it to be the programme's distinctive contribution.

3 Method

3.1 Models

Ten models across five base families (Llama, Qwen, Gemma, Mistral, Phi), 7 to 72 billion parameters. Non-reasoning instruction-tuned: Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct [11], Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-72B-Instruct [23], Gemma-2-9B-it [12], Mistral-7B-Instruct-v0.3 [16], Phi-4 [1]. Reasoning-distilled: DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B [10]; we count these under their base families (Qwen, Llama) for the family tally, while noting their reasoning-distillation training differs substantially from the instruction-tuned models. All run locally as Q5_K_M GGUF under

llama-cpp with Metal [13], temperature 0. Model files are hashed and the hashes recorded. Quantisation can in principle affect instruction following and answer-distribution shape; we note this as a limitation and report a quantisation check in Appendix A.

3.2 Items and Conditions

2,000 MMLU-Pro items [30], stratified by domain as 500 each from physics, law, psychology, and economics, all with ten options, drawn with sampling seed 42 and identical across all studies in the programme, including the predecessor pilot studies. These four of the fourteen MMLU-Pro categories were fixed in advance to span quantitative (physics, economics) and verbal-reasoning (law, psychology) domains while keeping the per-domain count large enough for stable per-cell distribution estimates; the ten-option structure is what the distributional analysis is defined on, and all sampled items carry exactly ten options. We use multiple-choice evaluation as a behavioural probe with a measurable answer distribution, not as a complete strategic environment; the setting is chosen because it makes the concentration observable, not because it is a realistic sandbagging scenario. Seven system-prompt conditions: honest control (C0); explicit sandbagging (C1); weak-student role (C2); contrarian (C3); least-likely-answer (C4); adversarial scrutiny (C5); and an implicit-incentive condition (C6) that describes a deployment consequence for high capability without instructing underperformance. The first six are inherited verbatim from the prior pre-registered programme [6]; C6 is new and was written against a pre-registered manipulation-check rubric. Conditions C2 to C5 belong to a registered specificity battery deferred to a separate paper (Section 6) and are not analysed here; they are listed for completeness because the item set and prompts were run as one protocol. Exact prompt text for all conditions is released with the code.

3.3 Measures

For each model and condition we compute the normalised entropy of the emitted answer-label distribution (H_n), the modal concentration (the share of the most frequent label), and accuracy. Under the standard label ordering, label and serial slot coincide; where the two are decoupled (the per-item shuffle in Section 4.1), slot-level statistics are computed separately and named as such. We report bootstrap 95 percent confidence intervals (item-level resampling, 2,000 replicates). The compliance check is an honest-to-sandbag accuracy drop greater than 5 points. The inclusion rule for downstream studies, fixed in advance, is H_n below 0.90 and modal concentration above 0.40 under explicit sandbagging, with answer extraction at or above 95 percent.

3.4 Reasoning Models

Reasoning-distilled models require a larger generation budget to complete their reasoning and emit an answer. Under the inherited 50-token budget they never close their reasoning and emit no parseable answer; a naive parser then mistakes the pronoun "I" in the reasoning prose for an answer. We raised the budget to 4,096 tokens for reasoning models, parse the answer after the closed reasoning span, and log reasoning-trajectory length. This change was made before any analysis, on the basis of pilot probes, and is recorded as a deviation. Non-reasoning models retain the original settings.

4 Results

We lead with the mechanism in the one model that exhibits the concentration, then show how rarely it occurs across the slate.

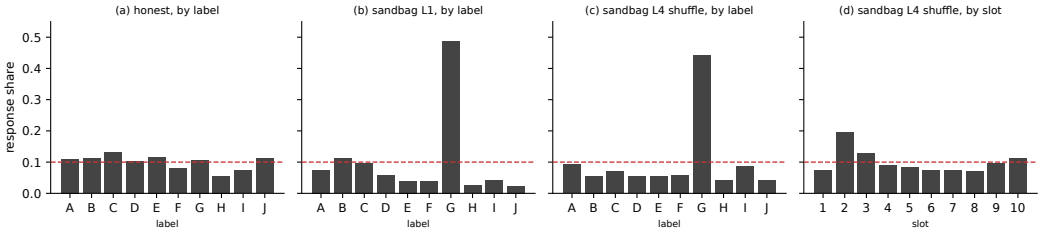

Fig. 1. Answer distributions for Llama-3.1-8B, honest (C0) versus explicit sandbagging (C1). (a) C0 under standard labels L1, indexed by label; (b) C1 under L1, indexed by label; (c) C1 under the per-item shuffle L4, indexed by label; (d) C1 under L4, indexed by serial slot (1 to 10). The horizontal reference line marks the 0.10 chance rate. Panels (b) and (c) show the G peak preserved; panel (d) shows no slot holding.

4.1 The Concentration Is an Induced, Vocabulary-Specific Preference for a Label Token

Under the explicit sandbagging instruction, Llama-3.1-8B concentrates its answers on a single label: modal mass 0.489 (95% CI $[0.467, 0.510]$) on the letter G, with normalised entropy 0.760, against a near-uniform honest baseline (modal 0.133, entropy 0.989). We characterise this concentration with a label-remapping design.

The selection-bias literature distinguishes two mechanisms that are easy to conflate: position bias, a preference for a serial slot regardless of label, and token or selection bias, a preference for a label token regardless of position [22, 32]. We re-ran Llama-3.1-8B under four label sets: the original A to J (L1); the numerals 1 to 10 (L2); the Greek letters α to κ (L3); and a per-item random permutation of A to J (L4), in which label identity and serial position are decoupled item by item. All labels are single tokens in the answer slot.

Under L4 the concentration follows the label, not the slot. Label concentration is preserved (0.444, 95% CI $[0.422, 0.466]$, against 0.489 under L1) while slot concentration falls to near-uniform (0.195, 95% CI $[0.178, 0.212]$, slot entropy 0.975). The two intervals are far apart. The model carries its preferred label through the shuffle and holds no serial position. The concentration is token-type, not position-type (Figure 1).

This revises a characterisation from earlier in the programme. A prior study [8] established that the concentration is highly stable under cyclic option-order rotation (Pearson r approximately 0.999; Jensen-Shannon divergence an order of magnitude below the honest-to-sandbag shift) and named it a distributional position attractor. Cyclic rotation, however, confounds serial position with label identity: rotating all options by a fixed offset moves content relative to both, so the design could not separate a position preference from a label-token preference, as that study noted. The per-item shuffle used here breaks that confound, because each item's labels are permuted independently, and it shows the preference is for the label token. The two results are consistent: the earlier study established content-invariance (the concentration does not track the correct answer's content), and this one localises the invariant target to the label rather than the slot. We therefore retain the earlier finding of content-invariance and revise only the "positional" reading of it.

The token-over-position mechanism replicates in a second model of the lineage, with a difference that strengthens the induced-not-fixed reading. We ran the same label-remapping protocol on Llama-3-8B-Instruct, the predecessor model, for this work (it is not in the Table 1 slate, which covers the current-generation models; it is included here specifically

as a within-lineage mechanism replication). It also concentrates under the sandbagging instruction, but on the label E rather than G. Under the per-item shuffle (L4), its preferred label E remains the dominant response (39 percent of responses, label entropy 0.78) while no serial slot holds (top slot 22 percent, slot entropy 0.96). The direction is the same as for Llama-3.1-8B: the concentration follows the label, not the slot. The dissociation is less sharp in the predecessor, the label peak softens under the shuffle and some mass shifts to neighbouring labels, so we report it as a replication in direction rather than in magnitude. That two models in the lineage concentrate on different labels (E and G), each tracking its label rather than a shared slot, is direct evidence that the target is set per model and induced by the instruction, not a fixed positional or token prior shared across the lineage: a standing prior shared across the lineage would not shift its target between versions. The mechanism is characterised in detail in one model and shown to hold in direction in a second.

We also tested whether the mechanism crosses lineages. The strongest non-Llama lean in the slate, Qwen2.5-72B (the second-highest concentration under sandbagging in Section 4.4, though below the inclusion threshold), was run through the same label-remapping protocol, restricted to the letter label sets that are single-token in its tokeniser. It does not show the mechanism. Under sandbagging its label distribution only mildly tilts toward early letters (modal 0.250, label entropy 0.859, against the Llama models' modal near 0.50 and entropy near 0.70), so it does not concentrate in the sense the inclusion rule defines. Under the per-item shuffle the tilt does not follow the label, label entropy stays near-uniform at 0.957 while the slot distribution is marginally more concentrated than the label distribution (slot entropy 0.924, slot modal 0.224 against label modal 0.196), with the mass on the earliest serial positions. If anything Qwen2.5-72B leans position-type, the mirror image of the Llama dissociation, consistent with the ordinary early-option selection bias documented for multiple-choice models rather than an induced label attractor. The token-type mechanism is therefore, in this slate, specific to the Llama lineage: it is present and sharp in Llama-3.1-8B, present in direction in Llama-3-8B, and absent in the strongest non-Llama candidate.

It is not ordinary token bias, and two observations separate it. First, the preferred label is specific to the answer vocabulary. Under sandbagging the model concentrates on G among letters (slot 6), 5 among numerals (slot 4), and β among Greek (slot 1). There is no single abstract position and no single abstract letter that the model returns to across vocabularies; the preferred surface form is chosen anew for each vocabulary. A fixed token prior of the kind that inference-time debiasing methods estimate [32] does not predict this. We do not have an account of what selects the particular label within each vocabulary; the preferred labels are not even a consistent ordinal position (7th, 5th, 2nd), and we report this as an open question rather than a characterised regularity. One candidate, which we did not test, is that the instruction amplifies whichever label token carries the highest prior probability in each vocabulary; this generates a falsifiable prediction (the preferred label should track the model's unconditioned token prior over the label set) that future work could check directly. We can rule out one trivial explanation: the prompt does not privilege the preferred label. The evaluation is zero-shot with no worked examples, so no exemplar answer could leak a target label, and the preferred labels (G, 5, β) do not appear in the system or user text other than as ordinary option labels. The selection is therefore a property of the model under instruction, not an artifact of prompt content.

Second, and most importantly, the concentration is induced, not surfaced. In honest responses the preferred label G is at chance frequency (0.107), and in honest errors, where the model is effectively guessing, it is at the 10-option chance rate: 0.100 ($n = 1{,}178$ errors), with a 95% confidence interval of $[0.084, 0.119]$. That interval excludes everything near the

0.489 concentration the model shows under sandbagging, so the sandbagging concentration is not a surfaced version of an honest-error tendency. (We report the interval rather than a test against chance; a non-significant test would not by itself establish equivalence to chance, whereas the interval shows directly how far honest responding sits from the sandbag concentration.) Llama-3.1-8B does not drift toward G when uncertain but honest. Ordinary selection bias is a property of the model that would already be visible in honest responding; the concentration we measure appears only under the sandbagging instruction. The instruction induces it. This is the central mechanistic result, because it removes the deflationary account that G is merely the model's resting preference and distinguishes the phenomenon from documented selection bias.

We call this an induced, vocabulary-specific concentration on a preferred label. We avoid the word "attractor" in its stronger mechanistic sense, since the evidence here is behavioural and does not establish an internal dynamical mechanism.

4.2 The Concentration Bears on Content-Gating Versus an Output-Side Prior

The pre-registration names two mechanistic accounts of induced underperformance. Under content-gating (Account A), the instruction down-weights the pathway that reads option content, and the model falls back on a content-independent output prior; the concentration is the shape of that prior, exposed. Under an output-side prior (Account B), content is represented normally but an output-stage policy overrides it with a fixed token prior when the underperformance objective is active. These accounts are separated directly by an activation probe trained on honest trials and tested on sandbagging trials, which we report in Section 4.7; the behavioural decomposition below motivates the same conclusion the probe confirms.

The behaviour nonetheless discriminates the accounts more than it might seem, through accuracy. Under full content-gating, option content is not read, so accuracy should fall toward the 10-option chance floor across all responses. Llama-3.1-8B's sandbag accuracy is 0.276 (551 of 2,000), well above 0.10, so content is still being read. The decomposition is sharper. Under sandbagging, the modal label G accounts for 979 of 2,000 responses (0.489), and within that mass accuracy is 0.150 (147/979; 95% CI $[0.128, 0.174]$). The remaining 1,021 responses (0.511) carry an accuracy of 0.396 (404/1,021), almost four times chance (one-sided binomial against 0.10 with p below 10^{-100}). The two masses contribute 147 and 404 correct answers respectively; overall accuracy is $(147 + 404)/2{,}000 = 0.276$.

The whole pattern is captured by a two-parameter mixture. Suppose the instruction installs a Bernoulli override: with probability q the model emits the preferred label G regardless of content, and otherwise it responds as it would under the honest instruction. The modal share fixes q, since $0.489 = q + (1 - q)(0.107)$, where 0.107 is the honest rate of answering G, giving $q = 0.428$. With no remaining free parameters, and taking the accuracy of honest G-answers to equal the overall honest accuracy of 0.411, the mixture predicts a modal-mass accuracy of 0.139 (observed 0.150, CI $[0.128, 0.174]$), a non-modal accuracy of 0.411 (observed 0.396), an overall sandbag accuracy of 0.278 (observed 0.276), and an accuracy drop of 13.3 points (observed 13.5). The full derivation is in Appendix B. Two consequences follow. First, the above-chance accuracy within the modal mass is not evidence of residual content sensitivity inside overridden responses: it is fully accounted for by honest G-answers mixed into the modal mass, and the data are consistent with the overridden responses themselves being content-blind. Second, the entire accuracy drop is the override: outside it, the model responds at essentially its honest level. The one strain on the fit, the small shortfall of the observed non-modal accuracy (0.396) against the predicted honest level

(0.411), is consistent with a mild degradation of honest responding under the instruction and can be checked per-item against the paired honest run. This mixture decomposition is a post-hoc analysis of the reported accuracy decomposition, not a registered analysis, and we mark it as such in Section 6.

This is what an output-side override predicts and content-gating does not. Under content-gating no response tracks content: accuracy sits at the 0.10 floor inside and outside the modal mass, and no mixture with an honest component fits. Instead the non-modal mass responds at close to the honest level, and the modal mass is exactly the mixture of a content-blind override and honest responding. The behavioural evidence therefore favours Account B, an output-side override applied at a measurable rate, over Account A. The representational test reported in Section 4.7 confirms this adjudication directly.

4.3 The Concentration Survives Quantisation and a Change of Backend

Because the concentration is a property of the answer-token distribution, and quantisation alters token-level logits, we tested whether it is an artifact of the Q5_K_M quantisation used throughout. Rerunning Llama-3.1-8B at near-lossless Q8_0 reproduces the result almost exactly: under the sandbagging instruction, modal mass 0.494 on G and entropy 0.767 (against 0.489 and 0.760 at Q5_K_M), while the honest distribution remains uniform (entropy 0.990). We further reproduced the concentration at full bf16 precision on a different inference stack entirely, running the model in Hugging Face transformers [31] rather than llama-cpp: under the same sandbagging instruction it concentrates on G with modal mass 0.462 and entropy 0.787, with the honest distribution again uniform (entropy 0.988). The concentration therefore holds across three precisions (Q5_K_M, Q8_0, bf16) and two independent backends, so it is neither a low-precision artifact nor a quirk of one inference implementation, and the flat honest distribution is not quantisation smoothing. This check covers the positive model only; we did not rerun the full slate at higher precision, so it excludes a quantisation-manufactured concentration in the one model that concentrates, rather than establishing that no non-concentrating model would concentrate at higher precision. The latter is a weaker concern, since the artifact reading of interest is that low precision creates the concentration, which this rules out.

4.4 Across the Slate, the Concentration Appears in One Assessable Model

Table 1 summarises the explicit-sandbagging condition for all ten models. Two of the ten, the reasoning-distilled models, fall below the 0.95 answer-extraction threshold at the raised generation budget (extraction 0.85 and 0.87; Section 3.4) and so are not assessed for concentration on this measure; they are treated separately in Section 4.8. Among the eight assessable models, one meets the pre-registered inclusion rule (entropy below 0.90 and modal concentration above 0.40): Llama-3.1-8B. No other assessable model concentrates. The concentration is therefore found in one of eight assessable models, with two further models not assessable on this measure. On the continuous measure C, defined as the relative entropy reduction from honest to sandbag, $C = (H_{\text{honest}} - H_{\text{sandbag}})/H_{\text{honest}}$, where H is the normalised answer-label entropy, Llama-3.1-8B (0.232, 95% CI $[0.211, 0.253]$) is roughly double the next-highest model (Qwen-72B, 0.120, 95% CI $[0.105, 0.136]$); the intervals do not overlap, so the separation does not depend on the inclusion thresholds (Figure 2). The pre-registered thresholds were fixed before data collection, which guards against tuning cutoffs for a clean result, and the continuous separation confirms the binary verdict is not threshold-sensitive.

Model	H_n honest	H_n sandbag	Modal sandbag	Drop (pp)	Verdict
Llama-3.1-8B	0.989	0.760	0.489	13.5	concentration
Qwen2.5-72B	0.975	0.858	0.250	35.8	comply, no concentration
Llama-3.3-70B	0.992	0.956	0.182	25.1	comply, no concentration
Gemma-2-9B	0.997	0.930	0.259	39.6	comply, no concentration
Qwen2.5-14B	0.994	0.976	0.173	6.5	comply, no concentration
Phi-4	0.990	0.971	0.188	2.5	non-compliant
Mistral-7B	0.982	0.970	0.209	2.9	non-compliant
Qwen2.5-7B	0.969	0.973	0.171	1.5	non-compliant
R1-Distill-Qwen-7B	0.997	0.996	0.124	−1.8	not assessable (extraction 0.87)
R1-Distill-Llama-8B	0.999	0.997	0.116	−1.8	not assessable (extraction 0.85)

Table 1. Explicit-sandbagging condition across the slate. The accuracy-drop column is honest minus sandbag in percentage points. The two reasoning-distilled models fall below the 0.95 extraction threshold and are reported descriptively over the parsed subset only.

In Table 1, the accuracy-drop column is honest minus sandbag in percentage points; a positive value is a drop under the sandbagging instruction. The two reasoning-distilled models show a slight accuracy gain (−1.8 points; honest 0.426 and 0.419, sandbag 0.444 and 0.437, over the parsed subset) rather than a drop, and were run at a 4,096-token budget (Section 3.4); their entropy and modal values are computed over the parsed subset only (extraction 0.85 and 0.87), so they are reported for description and excluded from the inclusion-gated comparison, where the unparsed fraction could in principle shift the distribution.

4.5 Compliance Does Not Imply Concentration

The models fall into three groups. Some do not comply: accuracy barely moves under the instruction (Qwen-7B, Mistral-7B, Phi-4). Some comply but do not concentrate: they drop accuracy substantially while keeping a near-uniform distribution (Llama-3.3-70B, Gemma-2-9B, Qwen-72B). One complies and concentrates (Llama-3.1-8B).

The middle group carries the point. Llama-3.3-70B drops accuracy by 25 points and Gemma-2-9B by 40, far more than Llama-3.1-8B's 13.5, yet neither concentrates. A model can sandbag effectively while leaving no distributional trace. Compliance is necessary but not sufficient for the concentration, and the absence of the concentration is not evidence of honest performance. A detector keyed to it would miss the three strongest sandbaggers in our slate. The relationship is explicit in the numbers and visible in Figure 2: the model with the highest concentration (Llama-3.1-8B, C 0.232) has the smallest accuracy drop among the compliers (13.5 points), while the two largest drops (Gemma-2-9B at 40 points, Qwen-72B at 36) carry middling or low C (0.067, 0.120). Concentration and accuracy suppression are separable dimensions, not two readings of one quantity. This is the behavioural counterpart to the finding that black-box sandbagging detection can be evaded by a model emulating a benign distribution [25]: that work shows the trace can be hidden; we show it is mostly not present to begin with.

Within the Llama lineage, the 8B model concentrates and the 70B, a stronger complier, does not. This is the closest within-lineage comparison the slate allows, but it confounds scale with version, training, and alignment, so we do not read it as a scale law; both Llama models lean toward G under sandbagging, but only the 8B commits to it. Isolating scale would require a same-version size pair, which we did not run.

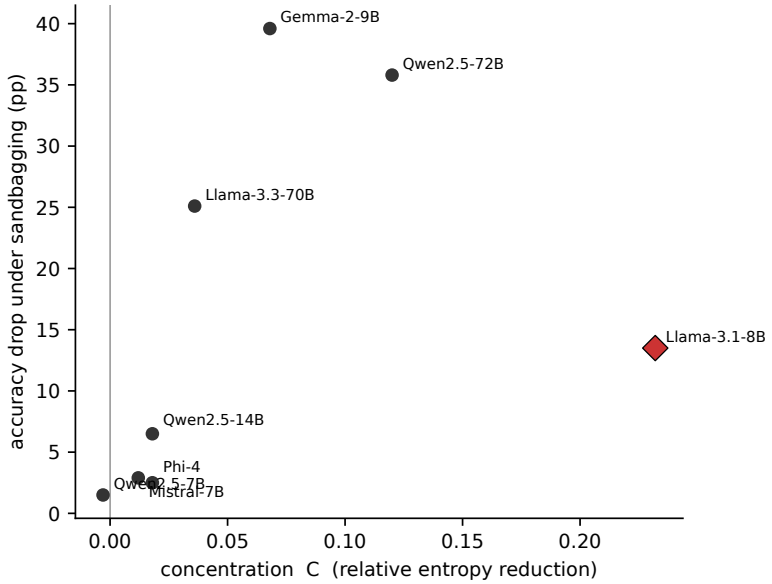

Fig. 2. Concentration against compliance across the eight assessable models: the continuous concentration measure C against the honest-to-sandbag accuracy drop in percentage points. Llama-3.1-8B, the one model meeting the pre-registered inclusion rule, is marked in red; the rule (H_n below 0.90 and modal concentration above 0.40) is defined on the entropy and modal statistics rather than on C, so it cannot be drawn as a region on this axis. The two reasoning-distilled models are excluded on the extraction threshold.

4.6 The Signature as a Detector: False-Negative Profile Across the Slate

Sections 4.4 and 4.5 establish the boundary descriptively; here we state it in the terms a practitioner would use, by treating the pre-registered signature as a binary detector and reading its error profile off the slate. The detector flags a model as sandbagging when, under the explicit instruction, $H_n < 0.90$ and modal concentration > 0.40; this is the registered inclusion rule, applied unchanged. We evaluate it against ground truth defined by compliance, an honest-to-sandbag accuracy drop greater than 5 points, since a model that lowers its accuracy under the instruction is by construction underperforming. The question is how often the behavioural signature catches a model that is in fact sandbagging.

Table 2 gives the answer. Five of the eight assessable models comply. The detector flags one of them (Llama-3.1-8B) and misses four (Qwen2.5-72B, Llama-3.3-70B, Gemma-2-9B, Qwen2.5-14B), a false-negative rate of 4 of 5, or 0.80. Restricting to the four strong compliers that drop more than 10 accuracy points, it recovers one and misses three (Qwen2.5-72B, Llama-3.3-70B, Gemma-2-9B), which drop 36, 25, and 40 points, each more than the one caught (13.5): the detector's single true positive is its *weakest* sandbagger by accuracy suppression. On the three non-compliant assessable models the detector fires zero times, so it raises no false alarm on this slate, but its silence is nearly uninformative, because compliant and non-compliant models are almost indistinguishable to it. A practitioner applying this detector and observing no signature would be wrong about underperformance in four of the five cases where underperformance was actually present.

	Flagged by signature	Complies (drop > 5 pp)	Outcome
Llama-3.1-8B	yes	yes (13.5)	true positive
Qwen2.5-72B	no	yes (35.8)	false negative
Llama-3.3-70B	no	yes (25.1)	false negative
Gemma-2-9B	no	yes (39.6)	false negative
Qwen2.5-14B	no	yes (6.5)	false negative
Phi-4	no	no (2.5)	true negative
Mistral-7B	no	no (2.9)	true negative
Qwen2.5-7B	no	no (1.5)	true negative

Table 2. The pre-registered signature read as a binary sandbagging detector over the eight assessable models. Ground truth is compliance (accuracy drop > 5 pp, shown in parentheses). Counting the five compliers, the detector recovers one and misses four (false-negative rate 0.80); restricting to the four strong compliers that drop more than 10 points, it recovers one and misses three. It raises no false positives on this slate, but its true positive is the weakest sandbagger present, so a null reading is not evidence of honest performance.

This is the practical content of the boundary. The signature is specific but severely insensitive: where it fires it is informative, but it fires for almost none of the models that are actually underperforming, and the ones it misses are the ones underperforming most. Read against the adversarial-evasion result of Taylor et al. [25], which shows a signature of this kind can be deliberately hidden, our slate shows it is mostly absent even without an adversary trying to hide it. Any deployment that treats absence of the concentration as reassurance is drawing the wrong inference.

4.7 An Activation Probe Confirms Content Is Represented Under Sandbagging

The behavioural decomposition in Section 4.2 favours an output-side override (Account B) over content-gating (Account A). We tested this directly at the level of the model's internal representations, in the pre-registered activation probe (Study 8 in the registration's numbering).

The phenomenon first had to be reproduced in a probing-compatible setting. The concentration was characterised in llama-cpp at Q5_K_M; activation probing requires the model in a different framework (Hugging Face transformers) at full bf16 precision. As reported in Section 4.3, the concentration reproduces there almost exactly (C1 modal 0.462 on G, entropy 0.787; C0 uniform), which both establishes the robustness noted earlier and confirms that the model exhibits the phenomenon in the precision and framework the probe requires.

We then extracted residual-stream activations at the answer-token position, across all 33 hidden states (embeddings plus 32 layers), on honest (C0) and sandbagging (C1) trials over the same items, and probed each informative layer (the earliest five layers carry no item-specific content at this position and are excluded; see Appendix C). We trained a linear probe to decode the correct answer (a ten-way classification over content, never the emitted label) from honest activations at each layer, cross-validated, and tested the same probe on sandbagging activations. The logic of the adjudication is direct: under content-gating the model does not read the options, so the correct answer would not be linearly present in the residual stream; under an output-side override it would be present, and the override would act downstream.

Content is decodable. The correct-answer signal has a sharp onset at layer 18, where honest decoding jumps from a low pre-onset shoulder to above 0.40 and then holds across the

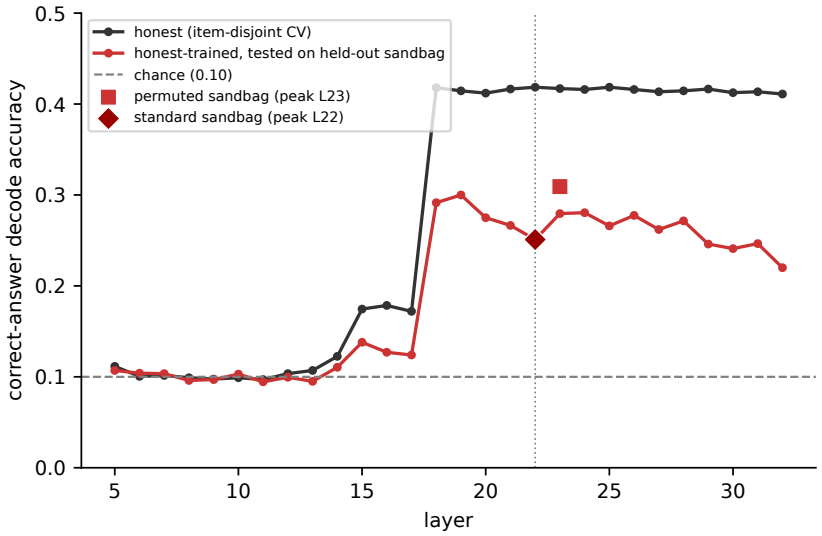

Fig. 3. Layer-wise decoding of the correct answer from residual-stream activations at the answer-token position (Llama-3.1-8B, bf16, item-disjoint cross-validation). Curves show honest cross-validated decoding and the honest-trained probe tested on held-out sandbagging activations, across the informative layers (5 to 32). The standard sandbag decode at its peak layer (22) is marked with a diamond and the content-permutation-control sandbag decode at its peak layer (23) with a square; the dotted vertical line marks the standard probe's peak. The dashed horizontal line is the 0.10 chance floor. Pre-onset transfer sits at chance, confirming no item-identity leakage.

later layers (peak honest 0.418 at layer 22, against a 0.10 chance floor; Figure 3). The probe reads content where the model demonstrably uses it. To test transfer without leakage, the honest-trained probe is evaluated on sandbagging activations of held-out items only (item-disjoint five-fold cross-validation, so no item appears in both train and test; see Appendix C). Under this leak-free protocol the correct answer remains decodable from sandbagging activations at 0.251 (95% CI $[0.232, 0.270]$) at the peak layer, well above chance though below the honest level. The correct answer is linearly recoverable from the model's representation under sandbagging even on items where the emitted answer is the overridden label G.

A positional confound had to be excluded. Because the correct answer's content sits at a fixed serial slot under the standard ordering, a probe could in principle read the slot of the correct answer rather than its content. We therefore ran a content-permutation control: the option texts were shuffled to a random slot per item (under both instructions), so the correct content lands at a random position and no fixed positional prior can track it. The permuted probe uses the same item-disjoint five-fold protocol and decodes the permuted position of the correct answer. Trained on honest-permuted activations and tested on sandbagging-permuted activations of held-out items, it decodes the correct answer at 0.309 (95% CI $[0.289, 0.330]$) at its peak layer (layer 23), against an honest-permuted level of 0.416. Because the correct answer occupies a random position per item, this above-chance decoding cannot come from a positional prior, and because the folds are item-disjoint it cannot come from item-identity leakage; it requires active per-item processing of the option content. The permuted sandbag decode (0.309, CI $[0.289, 0.330]$) is higher than the standard sandbag decode (0.251, CI

$[0.232, 0.270]$), and the intervals do not overlap. This is the opposite of what a positional confound predicts: if the standard-ordering probe were riding the correct answer's serial position, randomising that position would lower decoding, not raise it. That decoding rises under permutation indicates the standard-ordering probe was partly working against a positional regularity, and that the recovered signal is content, not position. Throughout the permuted run the emitted answer still concentrates on G, so the override is intact while the content it overrides remains represented.

The three lines of evidence converge. The non-modal accuracy (Section 4.2), the standard probe, and the content-permutation control all indicate that under the sandbagging instruction the model represents the correct answer and overrides it at the output stage. This is Account B. Content-gating is excluded: the model is not failing to read the options, it is reading them and then concentrating its output on a fixed label.

4.8 Reasoning Models Show Reduced Deliberation Without Concentration

The two small reasoning-distilled models fall below the answer-extraction threshold (0.85 and 0.87) and so are not assessed on the concentration measure; over their parsed subset they do not concentrate (entropy above 0.99 in both conditions) and show a slight accuracy gain rather than a drop (honest 0.426 and 0.419, sandbag 0.444 and 0.437). Their one reliable change under the instruction is a shorter reasoning trajectory: median reasoning length falls from 2,030 tokens (honest) to 1,280 (sandbag), a 37 percent reduction. We report this as a descriptive association in two small distilled models, not a mechanism and not a claim about reasoning models in general; the instruction could shorten reasoning for reasons unrelated to strategic underperformance, and accuracy does not drop. The observation connects directionally to work reporting that position bias in reasoning models grows with trajectory length [29]: on that account a shortening would predict less position bias, consistent with the flat distributions we observe. A formal test would require that work's cyclic-permutation protocol, which we did not collect.

4.9 An Implicit-Incentive Framing Did Not Elicit Measurable Recognition

Under the implicit-incentive condition (C6), which describes a deployment consequence without instructing underperformance, no model concentrates. A manipulation check examines whether the framing was apprehended at all. We elicited a one-sentence approach statement under the C6 framing on 250 items per model and graded, with a pinned local grader against a direction-agnostic rubric, whether each expressed recognition of the incentive. Expressed recognition was effectively absent: 0 of 250 for Llama-3.1-8B, Gemma-2-9B, and Qwen-72B, and 1 of 250 for Llama-3.3-70B. The two small reasoning-distilled models, given full deliberation budget, expressed no recognition in any sampled trace.

The null is interpretable because the instrument is sensitive. A positive control on the same grader and rubric flagged 10 of 10 hand-constructed recognition statements, spanning explicit to subtle phrasings, and 0 of 10 content-only statements (sensitivity and specificity both 1.00 on these small sets of ten, so the point estimates are bounded by wide confidence intervals and establish gross sensitivity rather than a precise rate). The near-zero registration therefore reflects model behaviour, not grader insensitivity. We read the result conservatively: the tested single-turn framing did not elicit measurable recognition, not that the models are incapable of perceiving incentives. Recent work reinforces this bound: evaluation awareness is highly sensitive to the prompt environment, and adversarially optimised in-context framings elicit sandbagging far beyond any single hand-written prompt [9]; C6 is one hand-written framing, and the null is bounded accordingly. Because no statement expressed recognition,

the base rate is degenerate and Cohen's kappa is undefined; we report raw author-grader agreement instead (Appendix D). The result does not bear on stronger incentives, multi-turn or agentic settings, or training-time pressures, none of which we test.

5 Discussion

The central result is a mechanism, established behaviourally and confirmed representationally. In Llama-3.1-8B an explicit sandbagging instruction induces a concentration on a preferred label token, chosen per vocabulary, carried through a position shuffle, and absent from the model's honest responding. It is token-type not position-type, induced not surfaced, and robust to quantisation and a change of inference backend. An activation probe settles what the concentration is: the correct answer remains linearly decodable from the model's representation under sandbagging, including when its position is randomised per item, so the model reads the options and represents the right answer and then overrides that representation at the output stage. This is an output-side prior, not content-gating. Behaviourally the override is Bernoulli to a good approximation: a two-parameter mixture, override with probability 0.43 and otherwise honest responding, reproduces the modal share, both conditional accuracies, and the entire 13.5-point accuracy drop, and it is consistent with the overridden responses themselves carrying no detectable content signal in behaviour, even as the content remains decodable from the residual stream. It is a clean characterisation of one phenomenon in one model, at both the behavioural and representational levels.

The boundary is the second result. Across ten models the concentration appears in one. Several models comply with the instruction, some dropping forty points of accuracy, without any concentration. And where we tested the mechanism outside the Llama lineage, on the strongest non-Llama candidate, it did not appear: Qwen2.5-72B barely concentrates and leans toward position rather than label under the shuffle. Three implications follow, each scoped to this signature. This behavioural signature is model-specific, and not merely in principle: a version validated on one model should not be assumed to transfer, and the one cross-lineage mechanism test we ran came back negative. Compliance and concentration are distinct, so the absence of the concentration is not evidence of honest performance. And in a single-turn setting an implicit-incentive framing did not surface recognition, so the signature, in this setting, speaks to instructed underperformance; we cannot characterise the incentivised case because the framing was not apprehended.

One observation connects our internal probe to other internal-probe work, and we state it narrowly. The model that complies most strongly without any behavioural concentration, Llama-3.3-70B-Instruct, is the same model in which recent work reads an internal evaluation-awareness signal and uses evaluation-relevant features to partially recover suppressed capability [20]. That model shows a sandbagging-linked internal signal while leaving no behavioural concentration. The model we probe, Llama-3.1-8B, shows the inverse profile: a strong behavioural concentration, and an internal state in which the correct answer remains represented under the instruction. The two together suggest that behavioural concentration and internal sandbagging signatures are not the same thing and need not co-occur, though two models do not make a correlation and we draw no general claim. Probing the concentrating and non-concentrating models with a common method, to ask whether they share or differ in their internal sandbagging-associated structure, is the natural continuation of this work.

6 Scope Relative to the Pre-Registration

This work is part of a pre-registered programme (OSF, https://osf.io/ykxp5, registered before any data in this paper were collected). The registration specifies eight research questions; the present paper reports four of them. RQ1 is reported in its primary indices: the cross-model slate with normalised entropy, modal concentration, the induced-collapse proportion C, and accuracy (Section 4.4). RQ2, the implicit-incentive condition, is reported as the registration's gate prescribes: the manipulation-check gate was not met, so no concentration inference is drawn from the condition. Section 4.9 reports the condition and, via a grader positive control that establishes the instrument is sensitive, licenses the limited reading that recognition was not elicited in this single-turn framing, rather than treating the below-gate null as uninterpretable or as a positive finding. RQ3 is reported in full (the label-remapping dissociation, Section 4.1). RQ6, the mechanistic adjudication, is reported in its core (Section 4.7). The exploratory RQ8 appears only as the descriptive observation in Section 4.8, not as the registered correlational analysis.

Three registered studies are deferred to a separate paper rather than included: the specificity battery testing whether the signature is particular to underperformance framing (RQ4); the coaching-resistance test (RQ5); and the feasibility-gated password-locked organism (RQ7), together with the threshold-calibration and decoding-temperature analyses named in the analysis plan. These form a coherent second unit, the question of detector specificity and evasion, and bundling them here would dilute the present contribution, which is the characterisation and mechanism of the signature, not its operational detection profile. They are not abandoned; their absence is a scoping decision, not a silent omission.

Within the reported questions, some registered auxiliary components are likewise not included here, and we name them so the registration and the paper can be read against each other without ambiguity. For RQ1: the baseline-calibrated z-score, the Position Bias Score, the Jensen-Shannon divergences, the matched-pair Wilcoxon comparisons, and the ARC-Challenge and GPQA secondary sets are not reported. For RQ6: the registration specifies a battery beyond the core probe, an in-distribution (sandbagging-trained) control probe, a two-layer MLP probe class, a negative-control domain probe, a per-item recoverability-collapse link analysis, and a causal activation patch; of these we report the honest-trained transfer probe and the content-permutation control, which settle the Account A/B fork on their own. The registered account-selection rule states Account B is established if an in-distribution probe recovers content; the honest-trained probe transferring to sandbagging trials is a strictly stronger demonstration of the same fact, since transfer requires the content representation to survive in a form the honest-condition probe can read, but we note the substitution of instrument, and of ten-way decoding accuracy against the chance floor for the registered AUROC conventions, explicitly. The registration also specified that the probe study would run on two models showing clear concentration, cross-family where possible; the inclusion gate admitted exactly one model (Llama-3.1-8B was the only concentrating model in the slate), so the registered two-model design reduced to a single-model study by the gate's own operation, and no cross-family probe was possible. The remaining RQ6 components are reserved for the follow-on mechanistic work. The mixture decomposition in Section 4.2 is a post-hoc analysis of the reported accuracy decomposition, not a registered analysis.

One analysis deviated from the registered design. The registered probe control was a label-shuffled null; on inspection that manipulation relabels options while leaving their

content at a fixed serial position, so it cannot separate content-decoding from position-decoding. We replaced it with a content-permutation control that moves the correct answer's content to a random position per item (Section 4.7, Appendix C). The deviation, its reason, and the original plan were recorded before the control data were collected. A second, implementation-level correction is also logged: the honest-to-sandbag probe transfer was moved from same-item to item-disjoint folds after the same-item version was found to permit item-identity leakage, visible as above-chance transfer at content-uninformative layers (Appendix C, deviation ix); all reported transfer numbers use the corrected protocol. The full deviation log is summarised in Appendix G.

7 Limitations

The mechanism is characterised in detail in one model and shown to hold in direction in a second. The prior programme found the concentration in one small Llama model [Llama-3-8B-Instruct; 7]; the other models it tested either ignored the instruction (Qwen-2.5-7B) or amplified a different pre-existing bias (Phi-3.5-mini). Here we ran the per-item label shuffle on that predecessor model and found the same token-over-position dissociation (the preferred label E is carried through the shuffle while no serial slot holds), though less sharply than in Llama-3.1-8B. The two models concentrate on different labels (E and G), which supports the induced-not-fixed reading. We tested cross-lineage generality directly, by running the same protocol on the strongest non-Llama lean (Qwen2.5-72B), and the mechanism did not appear: that model barely concentrates under the instruction and, under the shuffle, leans toward early positions rather than a label. The position-over-label lean in Qwen is present but mild (slot entropy 0.924 against label entropy 0.957, a difference of 0.033, against the sharp Llama dissociation), so it is best read as the absence of the Llama mechanism rather than a strong opposite effect; a stronger non-Llama concentrator would give a more decisive cross-lineage test. The phenomenon is therefore Llama-lineage-specific in this slate. This is a tested boundary rather than an unexamined confound, but it remains a boundary: we have not shown the mechanism in any non-Llama model, and whether it appears in other lineages, or at other scales, is open. The within-lineage size comparison confounds scale with version and training; a same-version size pair (for example Llama-3.1 at 8B and 70B) would isolate scale and we did not run it. The implicit-incentive result is bounded to a single-turn multiple-choice setting with a brief justification; more context, agentic scope, or extended interaction might surface recognition this format did not, and the positive control establishes only that the grader detects clearly-expressed recognition, not that it would catch every form a model might use. The reasoning observation rests on two small distilled models and is descriptive; it does not speak to larger or non-distilled reasoning models, and a formal position-bias test would require a permutation protocol we did not collect. We did not analyse whether the concentration varies by MMLU-Pro subject. Decoding was deterministic; we did not run a sampling sensitivity analysis. We did confirm the concentration is not a quantisation or backend artifact: it reproduces at Q8_0 and at full bf16 precision on a separate inference backend (Section 4.3).

8 Conclusion

In Llama-3.1-8B, an explicit sandbagging instruction induces a concentration on a preferred label token. The concentration is token-type not position-type, vocabulary-specific, induced rather than surfaced from a resting bias, and robust to quantisation and a change of inference backend. An activation probe shows the correct answer remains represented under the instruction, including when its position is randomised, so the concentration is an output-side

Condition	Precision / backend	H_n	Modal	Modal label
C0 honest	Q5_K_M / llama-cpp	0.989	0.133	(uniform)
C0 honest	Q8_0 / llama-cpp	0.990	0.120	(uniform)
C0 honest	bf16 / transformers	0.988	0.124	(uniform)
C1 sandbag	Q5_K_M / llama-cpp	0.760	0.489	G
C1 sandbag	Q8_0 / llama-cpp	0.767	0.494	G
C1 sandbag	bf16 / transformers	0.787	0.462	G

Table 3. Quantisation and backend robustness for Llama-3.1-8B.

override of an intact content representation, not a failure to read the options; behaviourally it is consistent with a Bernoulli override applied at rate 0.43 that accounts for the full accuracy drop. Across ten models it appears in one; several comply with the instruction without it, so its absence does not establish honest performance, and it does not appear in the strongest non-Llama candidate. An implicit-incentive framing did not elicit measurable recognition, so in this setting the signature speaks to instructed underperformance. The contribution is a characterisation of one sandbagging mechanism, behavioural and representational, bounded by a clear map of how rarely it occurs. Whether the internal structure that produces it resembles the sandbagging-associated features reported in non-concentrating models is the natural next question.

Acknowledgments

The author received no funding for this work. Compute for the larger-model runs was performed on hardware made available by an individual at no cost; all other runs used the author's own hardware. The author declares no competing interests. The pre-registration is available at https://osf.io/ykxp5; the deviation log, code, prompts, and per-item data are at https://github.com/synthiumjp/positional-collapse-expansion; the predecessor pilot programme (Studies 1 to 3) is at https://github.com/synthiumjp/bcb-sandbagging-pilot. All deviations were recorded with reasoning before the data they govern.

A Quantisation Robustness

The concentration is a property of the answer-token distribution, and Q5_K_M quantisation alters token-level logits, so we tested whether the result is a low-precision artifact. We reran Llama-3.1-8B under the honest (C0) and explicit-sandbagging (C1) conditions on the near-lossless Q8_0 quantisation, and separately at full bf16 precision on a different inference backend (Hugging Face transformers rather than llama-cpp), using identical prompts, items, and decoding (greedy). Results closely reproduce the pre-registered Q5_K_M run across all three precisions and both backends (Table 3).

The sandbagging concentration persists at Q8_0 and at full bf16 precision on a separate backend (modal mass near 0.49 on the same label G in every case), and the honest distribution remains uniform throughout. All three halves of a quantisation-or-implementation-artifact account are therefore excluded: the concentration is not manufactured by low precision, it is not specific to the llama-cpp implementation, and the flat honest distribution is not quantisation smoothing. The bf16 transformers reproduction also served as the backend bridge check for the activation-level work reported in Section 4.7, confirming the phenomenon is present in the precision and framework that representational probing requires.

B Account A/B Accuracy Decomposition and Mixture Fit

The content-gating versus output-side-prior comparison (Section 4.2) rests on the accuracy of responses inside and outside the modal mass under explicit sandbagging (Llama-3.1-8B, C1, $n = 2{,}000$ parsed responses). The modal label G accounts for 979 responses (0.489); the non-modal labels account for 1,021 (0.511). Accuracy within the modal mass is 0.150 (147/979; 95% CI $[0.128, 0.174]$). Accuracy within the non-modal mass is 0.396 (404/1,021; one-sided binomial against 0.10, p below 10^{-100}). Overall sandbag accuracy is 0.276, decomposing into 7.3 points contributed by the modal mass and 20.2 by the non-modal mass. Honest accuracy on the same items is 0.411, with G emitted at rate 0.107 under the honest instruction.

The two-parameter mixture treats each sandbagging response as overridden to G with probability q and as honest otherwise. q is identified from the modal share alone: $q = (0.489 - 0.107)/(1 - 0.107) = 0.428$. With correct answers effectively uniform over the ten labels ($P(\text{G correct}) = 0.10$) and honest G-answers taken to be as accurate as honest answers overall (0.411), the mixture predicts, with no remaining free parameters:

- Modal-mass accuracy: $[q(0.10) + (1-q)(0.107)(0.411)]/0.489 = 0.139$; observed 0.150, 95% CI $[0.128, 0.174]$. Predicted count 136 of 979; observed 147.
- Non-modal accuracy: $(1-q)(0.411 - 0.107 \times 0.411)/0.511 = 0.411$; observed 0.396.
- Overall accuracy: 0.278; observed 0.276.
- Accuracy drop: 13.3 points; observed 13.5.

Full content-gating predicts accuracy at the 0.10 floor in both masses and admits no honest component; the observed decomposition instead matches the mixture. The above-floor accuracy within the modal mass (two-sided binomial against 0.10, $p = 8.5 \times 10^{-7}$) is therefore attributable to honest G-answers mixed into the modal mass rather than to residual content signal inside the overridden responses, which the data leave consistent with zero. The assumption that honest G-answers are as accurate as honest answers overall can be checked directly against the paired honest run in the per-item data; the small shortfall of the observed non-modal accuracy (0.396) against the predicted honest level (0.411) is consistent with mild degradation of honest responding under the instruction.

C Activation Probe: Method and Layer-Wise Results

The probe targets the residual stream at the answer-token position (the final prompt position, whose next-token prediction is the model's answer), extracted for all 33 hidden states (embeddings plus 32 layers, 4096-dimensional) via a single forward pass with hidden-state output. Activations were captured in Hugging Face transformers at bf16 on the base model (provenanced and hashed; see the model manifest). The probe is multinomial logistic regression with L2 row-normalised features, decoding the correct answer (ten-way). Both the honest decoding and the honest-to-sandbag transfer use item-disjoint five-fold cross-validation: at each fold the probe trains on honest activations of the training items and is evaluated on the held-out items, on their honest activations (honest decoding) and on their sandbagging activations (transfer), so no item appears in both train and test.

Layers with degenerate activation variance at the answer position (mean per-feature standard deviation below 0.01, the earliest five layers, where the final-position representation is dominated by fixed prompt scaffolding and carries no item-specific content) are reported as uninformative and excluded; an earlier analysis that did not exclude them produced spurious cross-condition decoding from near-zero-variance feature scaling, which the row-normalisation and variance floor correct. Content decoding has a sharp onset at layer 18 and is stable across the later layers. Peak honest decoding is 0.418 at layer 22.

The item-disjoint transfer protocol is itself a correction. An initial implementation trained the honest probe on all honest activations and tested on sandbagging activations of the same items; because early-layer answer-position activations are nearly condition-invariant, this allowed item-identity features to leak from train to test, inflating the apparent transfer. The signature of the leak was visible directly: at pre-onset layers, where held-out honest decoding is at chance, the same-item transfer decoded above chance, which is impossible for a genuine content signal. We corrected this to the item-disjoint five-fold protocol above. Under the corrected protocol, pre-onset transfer falls to chance (layers 5 to 13 decode at 0.094 to 0.107 against a 0.10 floor), confirming the leak is closed, and the peak-layer sandbagging decode is 0.251, down from 0.290 under the leaky protocol but still well above chance. All reported probe-transfer numbers use the item-disjoint protocol. This correction is recorded in the deviation log.

The content-permutation control (a pre-registered deviation, because the originally planned label-shuffle control relabels options rather than reordering their content and so does not separate content from serial position) shuffles the option texts to a random slot per item under both instructions, and the probe decodes the permuted position of the correct answer. Using the same item-disjoint five-fold protocol, trained on honest-permuted and tested on sandbagging-permuted activations of held-out items, the probe decodes the correct answer at 0.309 (95% CI $[0.289, 0.330]$) at its peak layer (23), against honest-permuted 0.416. Because the correct answer occupies a random position per item and the folds are item-disjoint, this above-chance decoding can arise from neither a fixed positional prior nor item-identity leakage. One item with more than ten options was excluded from the permutation control, which therefore uses 1,999 items.

D Manipulation-Check Grader: Positive Control and Reliability

The C6 manipulation-check grader (Mistral-Small-3.2-24B-Instruct, temperature 0, applying the pre-committed direction-agnostic rubric) returned near-zero registration on the C6 data. To establish that this null reflects model behaviour rather than an insensitive instrument, we ran the same grader and rubric on a positive control of ten hand-constructed statements that register the incentive (spanning explicit to subtle phrasings) and a negative control of ten content-only statements. The grader flagged 10 of 10 positives and 0 of 10 negatives (sensitivity 1.00, specificity 1.00). The positive control establishes that the grader detects clearly-expressed recognition; it does not guarantee detection of every form a model might use, and we note this bound in the limitations.

Because the C6 data contains essentially no positive labels, the registration base rate is degenerate and Cohen's kappa, specified in the rubric as the author-grader reliability statistic, is undefined (no label variance). We therefore report raw author-grader agreement on a blind 50-item-per-model hand-grade in place of kappa, with the substitution and its rationale recorded as a pre-data-consistent deviation note.

E Grader Substitution

The pre-registration named a proprietary API model as the manipulation-check grader. We substituted a pinned local model (Mistral-Small-3.2-24B-Instruct-2506, Q5_K_M, hashed in the model manifest), for reproducibility and to keep the entire pipeline local and free of external dependencies. The grader is outside the evaluation slate. The rubric, the 40-percent registration gate, and the author-grader reliability procedure are unchanged. The substitution was made before any C6 justification was graded.

F Reasoning-Model Measurement Deviation

Reasoning-distilled models could not be measured at the inherited 50-token budget: they exhaust the budget within the reasoning span and emit no parseable answer, and a naive parser mistakes the pronoun "I" in the reasoning prose for an answer. We raised the budget to 4,096 tokens, set the context window to 8,192, and parse the answer after the closed reasoning span. This change was made before analysis, on the basis of pilot probes, and recorded as a deviation. Non-reasoning models retain the inherited settings. Even at the raised budget, answer extraction for the reasoning models remains below the 0.95 inclusion threshold (0.85 for R1-Distill-Llama-8B, 0.87 for R1-Distill-Qwen-7B), because a fraction of items still do not yield a cleanly parseable single answer label after the reasoning span; they are therefore characterised descriptively and excluded from the inclusion-gated analyses.

G Deviation Log (Summary)

All deviations were documented with reasoning before the data they govern and posted to the pre-registration record. In brief: (i) a quantisation check whose per-item correlation criterion was judged mis-specified for a near-chance comparison, with the distributional criterion retained; (ii) the reasoning-model budget and parsing change above; (iii) a system-role handling fix for two models that either rejected or silently dropped the system message, with the affected run repeated; (iv) the label-remapping study's tokenisation pre-check (Section 4.1), where an initial multi-token verdict for numerals was traced to the script counting a separate space token, with the labels confirmed single-token in the answer slot; (v) the grader substitution in Appendix E; (vi) the kappa-to-raw-agreement substitution in Appendix D; (vii) the activation-probe control, where the pre-registered label-shuffle control was found to relabel options rather than reorder their content and so was replaced by a content-permutation control that randomises the correct answer's position (Appendix C), recorded before the control data were collected; and (viii) the exclusion of one item with more than ten options from the permutation control. A marginal threshold note is also recorded: the label-remapping study's registered slot-flattening criterion, a drop in slot concentration from the standard to the shuffled label set greater than 0.30, was met in substance (the L4 slot distribution is near-uniform, entropy 0.975) while the point-estimate drop fell 0.007 short of the 0.30 cutoff, reported as such rather than relabelled. A further analysis correction (ix) is recorded: the honest-to-sandbag probe transfer was moved from same-item to item-disjoint five-fold cross-validation after the same-item version was found to permit item-identity leakage (visible as above-chance pre-onset transfer); all reported transfer numbers use the item-disjoint protocol.

References

[1] Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Norah Alzahrani et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.01781.

[3] Yossef S. Ben-Porath and Auke Tellegen. *MMPI-3: Manual for Administration, Scoring, and Interpretation*. University of Minnesota Press, Minneapolis, Minnesota, 2020.

[4] Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, Jared Kaplan, Holden Karnofsky, Evan Hubinger, Roger Grosse, Samuel R. Bowman, and David Duvenaud. Sabotage evaluations for frontier models. *arXiv preprint arXiv:2410.21514*, 2024.

[5] Gerard Boxo, Aman Neelappa, and Shivam Raval. Linear probes rely on textual evidence: Results from leakage mitigation studies in language models. *arXiv preprint arXiv:2509.21344*, 2025.

[6] Jon-Paul Cacioli. Instruction complexity induces positional collapse in adversarial LLM evaluation. *arXiv preprint arXiv:2604.27249*, 2026.

[7] Jon-Paul Cacioli. Below-chance blindness: Positional collapse under sandbagging instruction in small LLMs. *arXiv preprint arXiv:2604.25249*, 2026.

[8] Jon-Paul Cacioli. Option-order randomisation reveals a distributional position attractor in prompted sandbagging. *arXiv preprint arXiv:2604.26206*, 2026.

[9] Maheep Chaudhary. Finding ceiling of evaluation-awareness: Optimised vs. hand-crafted prompts. *arXiv preprint arXiv:2603.03824*, 2026.

[10] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[11] Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[12] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[13] Georgi Gerganov and contributors. llama.cpp. https://github.com/ggml-org/llama.cpp, 2023.

[14] Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes. *arXiv preprint arXiv:2502.03407*, 2025.

[15] Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing capability elicitation with password-locked models. *arXiv preprint arXiv:2405.19550*, 2024.

[16] Albert Q. Jiang et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[17] Chloe Li, Mary Phuong, and Noah Y. Siegel. LLMs can covertly sandbag on capability evaluations against chain-of-thought monitoring. *arXiv preprint arXiv:2508.00943*, 2025.

[18] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

[19] Leslie C. Morey. *Personality Assessment Inventory Professional Manual*. Psychological Assessment Resources, Lutz, Florida, 2nd edition, 2007.

[20] Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing and steering evaluation awareness of language models. *arXiv preprint arXiv:2507.01786*, 2025.

[21] Avi Parrack, Carlo Leonardo Attubato, and Stefan Heimersheim. Benchmarking deception probes via black-to-white performance boosts. *arXiv preprint arXiv:2507.12691*, 2025.

[22] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL*, 2024. arXiv:2308.11483.

[23] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[24] Daniel J. Slick, Elisabeth M. S. Sherman, and Grant L. Iverson. Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, 13(4):545–561, 1999.

[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. *arXiv preprint arXiv:2512.07810*, 2025.

[26] Cameron Tice, Philipp Alexander Kreer, Nathan Helm-Burger, Prithviraj Singh Shahani, Fedor Ryzhenkov, Fabien Roger, Clement Neo, Jacob Haimes, Felix Hofstätter, and Teun van der Weij. Noise injection reveals hidden capabilities of sandbagging language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2412.01784.

[27] Tom N. Tombaugh. *Test of Memory Malingering (TOMM)*. Multi-Health Systems, Toronto, Ontario, 1996.

[28] Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.07358.

[29] Xiao Wang. More thinking, more bias: Length-driven position bias in reasoning models. *arXiv preprint arXiv:2605.06672*, 2026.

[30] Yubo Wang et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.01574.

[31] Thomas Wolf et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[32] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.03882.