Does an honest-error probe detect deliberate sandbagging, or is deliberate wrongness internally distinct?

Abstract

Linear probes on residual activations can read whether a model is about to answer incorrectly [@alain2016probes] [@azaria2023internal] [@marks2024geometry], and internal representations can encode truth even when outputs do not [@burns2023ccs]. We ask whether an error-awareness probe trained only on a model's honest mistakes also fires when the model is deliberately sandbagging, or whether deliberate wrongness occupies a distinct internal direction. Across three instruct models, Llama 3.1 8B [@grattafiori2024llama3], Qwen2.5 7B [@qwen2024qwen25] and Mistral 7B [@jiang2023mistral], on 600 MMLU questions [@hendrycks2021mmlu] answered under an honest condition and a sandbag condition, we train an honest-error probe (HC versus HI) and test transfer to deliberate wrong answers (SC), with controls that hold the sandbag prompt constant. The honest-error probe appears to transfer (pooled HC versus SC AUC 0.8849), and at a honest-fixed threshold it flags sandbagging about twice as often as honest mistakes (pooled recall ratio 2.1836). That transfer is largely an artifact of the prompt: the same probe fires almost as hard on refusers (SN, sandbag prompt but correct answer) as on compliers, and once the prompt is held constant the honest-error direction barely separates deliberate wrong from refused correct (AUC 0.5280 to 0.5911). A probe trained directly within the sandbag condition separates deliberate wrong from refused correct only weakly (AUC 0.6646 to 0.7171), and that residual does not isolate intent. The same direction reads honest answer correctness about as well, and it vanishes on items the model could not answer, so it is consistent with a generic answer-correctness direction rather than a sandbag-specific one. We did not find a direction clean of both the prompt and answer correctness. The practical implication is that an error-awareness probe is not, on its own, a sandbagging detector [@park2024deception] [@pacchiardi2024liar].

Setup

For each model every MMLU question is run twice: an honest condition (answer to the best of ability) and a sandbag condition (deliberately give a wrong but plausible letter). Answering is direct, a single letter with no chain of thought [@turpin2023saywhat] [@lanham2023faithfulness], so the final pre-answer token carries the committed decision; we read the residual stream at that token for seven layers. Four groups follow: honest correct (HC), honest incorrect (HI, honest mistakes), sandbag complied (SC, the deliberate wrong answers), and sandbag noncomply (SN, refused or failed to sandbag, the same prompt as SC but a correct answer). All probes are L2 logistic regressions on standardized features, seed 42, with item-disjoint stratified k-fold so the same question never spans train and test; the layer is chosen by cross-validation on training data only [@belinkov2022probing].

Probe metrics

model	probe	task	layer	AUC	ci_lo	ci_hi	Cohen d
llama-3.1-8b-instruct	A_indist	HC_vs_HI	24	0.7348	0.6943	0.7723	0.7261
llama-3.1-8b-instruct	B_distinct	HI_vs_SC	8	1.0000	1.0000	1.0000	6339.0752
llama-3.1-8b-instruct	P_prompt	honest_vs_sandbag	8	1.0000	1.0000	1.0000	na
llama-3.1-8b-instruct	W_withinsb	SC_vs_SN	16	0.6646	0.6204	0.7081	0.5874
qwen2.5-7b-instruct	A_indist	HC_vs_HI	28	0.7611	0.7234	0.7989	0.8433
qwen2.5-7b-instruct	B_distinct	HI_vs_SC	7	1.0000	1.0000	1.0000	782.4294
qwen2.5-7b-instruct	P_prompt	honest_vs_sandbag	7	1.0000	1.0000	1.0000	na
qwen2.5-7b-instruct	W_withinsb	SC_vs_SN	24	0.7171	0.6767	0.7581	0.7107
mistral-7b-instruct-v0.3	A_indist	HC_vs_HI	12	0.7150	0.6764	0.7546	0.7649
mistral-7b-instruct-v0.3	B_distinct	HI_vs_SC	8	1.0000	1.0000	1.0000	47.8525
mistral-7b-instruct-v0.3	P_prompt	honest_vs_sandbag	8	1.0000	1.0000	1.0000	na
mistral-7b-instruct-v0.3	W_withinsb	SC_vs_SN	20	0.7063	0.6622	0.7475	0.7814

model	probe	task	layer	AUC	ci_lo	ci_hi	Cohen d
POOLED	A_indist	HC_vs_HI	pooled	0.7420	0.7191	0.7643	0.8037
POOLED	B_distinct	HI_vs_SC	pooled	1.0000	1.0000	1.0000	79.6482

Probe A (HC versus HI) reaches in-distribution AUC 0.7348, 0.7611 and 0.7150 on the three models (pooled 0.7420), the expected error-awareness range [@kadavath2022know] [@hendrycks2017baseline]. Probe B (HI versus SC) is perfectly separable (AUC 1.0000 everywhere), and the prompt detector P (honest versus sandbag) is also perfect (AUC 1.0000): the sandbag instruction is trivially decodable, so Probe B's perfect score is suspect. The within-sandbag probe W (SC versus SN) separates only moderately (0.6646, 0.7171, 0.7063).

Transfer of the honest-error probe to sandbagging

model	transfer AUC (HC vs SC)	ci_lo	ci_hi	naive AUC	Cohen d	honest thr	HI recall	SC recall	SC/HI
llama-3.1-8b-instruct	0.8721	0.8460	0.8974	0.9933	1.6669	0.9811	0.2555	0.5646	2.2098
qwen2.5-7b-instruct	0.8889	0.8644	0.9106	0.9897	1.8371	0.9397	0.3262	0.6417	1.9671
mistral-7b-instruct-v0.3	0.9026	0.8802	0.9240	0.9998	1.6754	0.9941	0.3077	0.7322	2.3796
POOLED	0.8849	0.8708	0.8985	na	1.7309	na	0.2957	0.6457	2.1836

Taken at face value the honest-error probe transfers: item-disjoint HC versus SC AUC is 0.8721, 0.8889 and 0.9026 (pooled 0.8849), higher than the in-distribution honest task, and at a threshold fixed to a 10 percent honest false-positive rate the probe flags 0.5646, 0.6417 and 0.7322 of sandbagging versus 0.2555, 0.3262 and 0.3077 of honest mistakes, a recall ratio of 2.2098, 1.9671 and 2.3796 (pooled 2.1836). A detector trained never having seen sandbagging fires on it more than on the errors it was trained on.

The transfer is mostly the prompt

model	prompt P AUC	within-sb W AUC	W Cohen d	A: SC vs SN (prompt-ctrl)	A mean HC	A mean HI	A mean SC	A mean SN	B mean HI	B mean SC	B mean SN	SN resembles SC
llama-3.1-8b-instruct	1.0000	0.6646	0.5874	0.5280	0.2695	0.5584	0.8394	0.7981	0.0000	1.0000	1.0000	1.0000
qwen2.5-7b-instruct	1.0000	0.7171	0.7107	0.5540	0.2138	0.5308	0.8233	0.7275	0.0000	0.9999	0.9998	1.0000

model	prompt P AUC	within-sb W AUC	W Cohen d	A: SC vs SN (prompt-ctrl)	A mean HC	A mean HI	A mean SC	A mean SN	B mean HI	B mean SC	B mean SN	SN resembles SC
mistral-7b-instruct-v0.3	1.0000	0.7063	0.7814	0.5911	0.3225	0.6350	0.8909	0.8074	0.0019	1.0000	1.0000	1.0000

The controls dismantle the transfer. The honest-error probe assigns mean incorrectness 0.2695, 0.2138 and 0.3225 to HC but 0.7981, 0.7275 and 0.8074 to SN, the group that received the sandbag prompt yet answered correctly. It fires on the prompt, not the wrong answer. Held within the sandbag condition, the honest-error direction separates deliberate wrong (SC) from refused correct (SN) at only 0.5280, 0.5540 and 0.5911, near chance. Probe B tells the same story from the other side: the refusers SN score 1.0000, 0.9998 and 1.0000 under Probe B, sitting fully on the SC side (sn_resembles_sc 1.0000), so Probe B's perfect HI versus SC separation is reading the sandbag prompt [@sharma2023sycophancy] [@chwang2024androids], not the intent. What survives the prompt control is the within-sandbag probe W at 0.6646, 0.7171 and 0.7063, a weak separation of deliberate wrong from refused correct. The next section asks whether that residual is sandbag-specific or merely a generic answer-correctness direction [@zou2023repe] [@li2023iti].

Is the within-sandbag signal sandbag-specific or generic correctness?

model	W to honest (HC vs HI)	A to sandbag (SC vs SN)	A SC vs SN (common layer)	A HC vs HI (common layer)	cos(A, W)	W honest-correct items	W honest-wrong items	W residualized
llama-3.1-8b-instruct	0.6711	0.5280	0.5200	0.6720	0.1212	0.6503	0.4454	0.6542
qwen2.5-7b-instruct	0.7389	0.5540	0.5530	0.7332	0.0621	0.8248	0.5180	0.6920
mistral-7b-instruct-v0.3	0.6665	0.5911	0.6169	0.6931	0.2703	0.7257	0.5355	0.6686

The within-sandbag direction W separates deliberate wrong answers (SC) from refused-correct answers (SN), so it confounds intent with plain answer correctness, and three tests show the separation is largely the correctness component. First, the W direction applied across context to honest answers separates honest correct from honest incorrect at AUC 0.6711, 0.7389 and 0.6665 (pooled 0.6437), about as well as it separates SC from SN in its own condition (0.6646, 0.7171, 0.7063) and about as well as the dedicated honest-error probe reads honest correctness at the same layer (0.6720, 0.7332, 0.6931). A direction that reads honest correctness as well as the honest-error probe does is a generic correctness direction, not a sandbag-specific one, and symmetrically the honest-error direction taken at that same layer barely reads the within-sandbag contrast (0.5200, 0.5530, 0.6169), close to its 0.5280, 0.5540 and 0.5911 at its own layer (pooled 0.5570). Second, stratifying SC versus SN by whether the model answered the same item correctly in the honest condition removes item difficulty, and the within-sandbag separation lives on items the model

could answer honestly (AUC 0.6503, 0.8248, 0.7257) while it vanishes on items it could not (0.4454, 0.5180, 0.5355, all consistent with chance). The signal is present only where there was a correct answer to abandon. Third, residualizing out the honest-error projection leaves the SC versus SN separation almost unchanged (0.6542, 0.6920, 0.6686), which shows W is not the same vector as the honest-error direction (cosine 0.1212, 0.0621, 0.2703) yet does not rescue an intent reading, since the surviving component still tracks answer correctness. We did not find a clean direction independent of both the prompt and answer correctness. The only prompt-independent signal is consistent with a generic wrongness direction [@marks2024geometry] [@azaria2023internal].

Figures

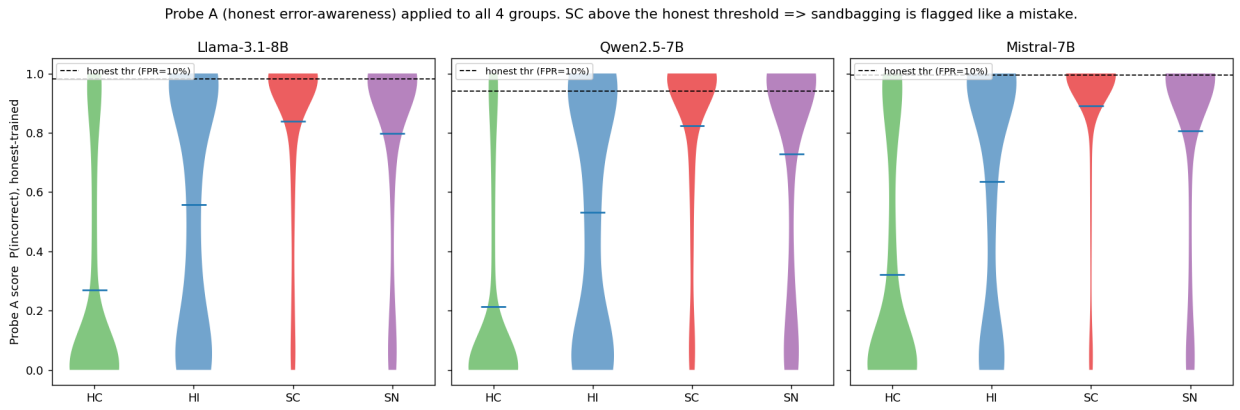

Figure 1: Probe A scores across the four groups, with the honest threshold

Answer

An error-awareness probe trained only on honest mistakes does not, by itself, detect deliberate sandbagging. Its apparent transfer (AUC near 0.88, recall ratio above 2) is an artifact of the sandbag instruction being present in the activations, since the probe fires almost as hard on correct-but-prompted answers as on deliberate wrong ones, and the prompt-controlled separation stays near chance (pooled 0.5570). The one signal that survives the prompt control, a probe trained within the sandbag condition (AUC about 0.66 to 0.72), does not isolate intent. It reads honest answer correctness about as well as it reads within-sandbag wrongness and it disappears on items the model could not answer, so it is consistent with a generic answer-correctness direction rather than a signature of deliberate wrongness. We did not find a direction clean of both the prompt and answer correctness. Measuring a prompt-independent intent signal needs a prompt-matched sandbagger such as a password-locked or fine-tuned model [@xiong2023canllms] [@tian2023justask] [@guo2017calibration] [@lin2022teaching].

Limitations

The central limitation is that the surviving within-sandbag contrast confounds intent with answer correctness. SC and SN share the sandbag prompt but differ in whether the committed answer is wrong, and they are not a matched pair, since SN is the endogenously selected subset of sandbag-prompted items the model answered correctly anyway. The disentangling tests show the separation tracks correctness rather than deliberateness, so the within-sandbag direction should be read as a within-sandbag error-awareness direction and not as a detector of intent.

Sandbagging here is elicited by an in-context instruction rather than arising naturally, and that instruction is trivially decodable from the activations (prompt AUC 1.0000). The prompt-in-activations confound is therefore intrinsic to this design, and every comparison that does not hold the prompt constant is inflated by it. The deployment-relevant case, a model that sandbags with no instruction in context, cannot be measured

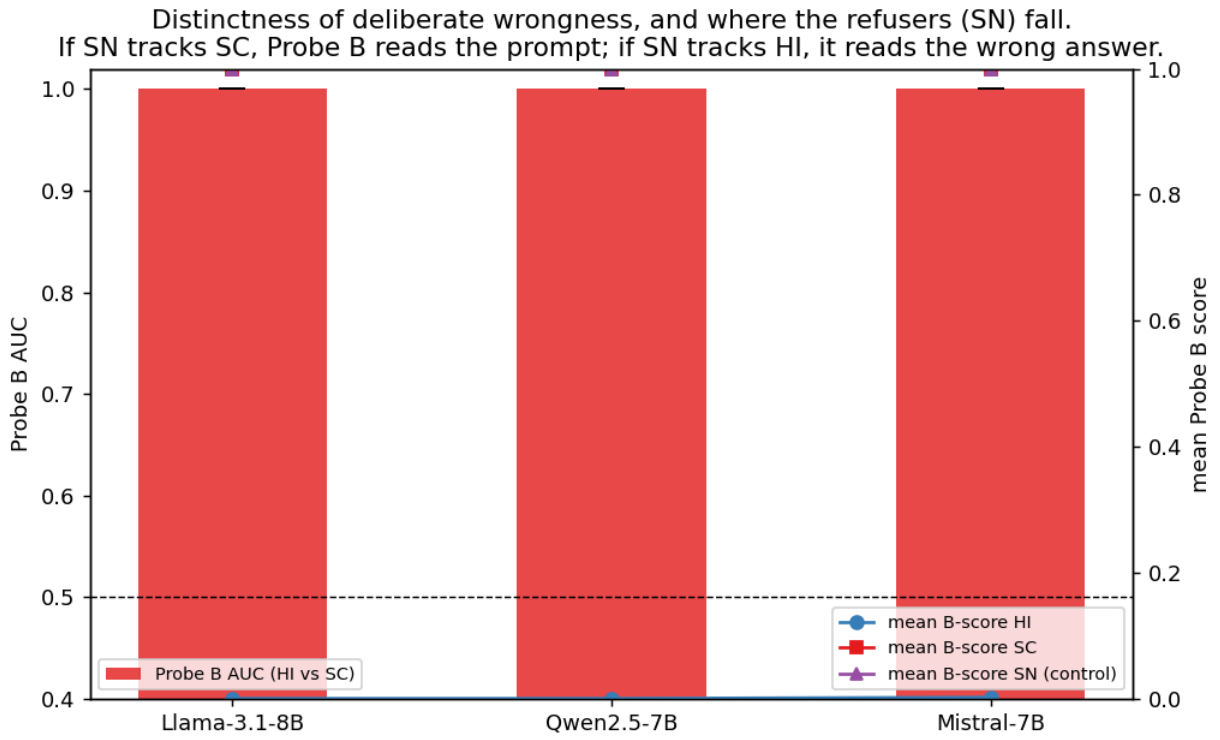

Figure 2: Distinctness of deliberate wrongness and where the refusers fall

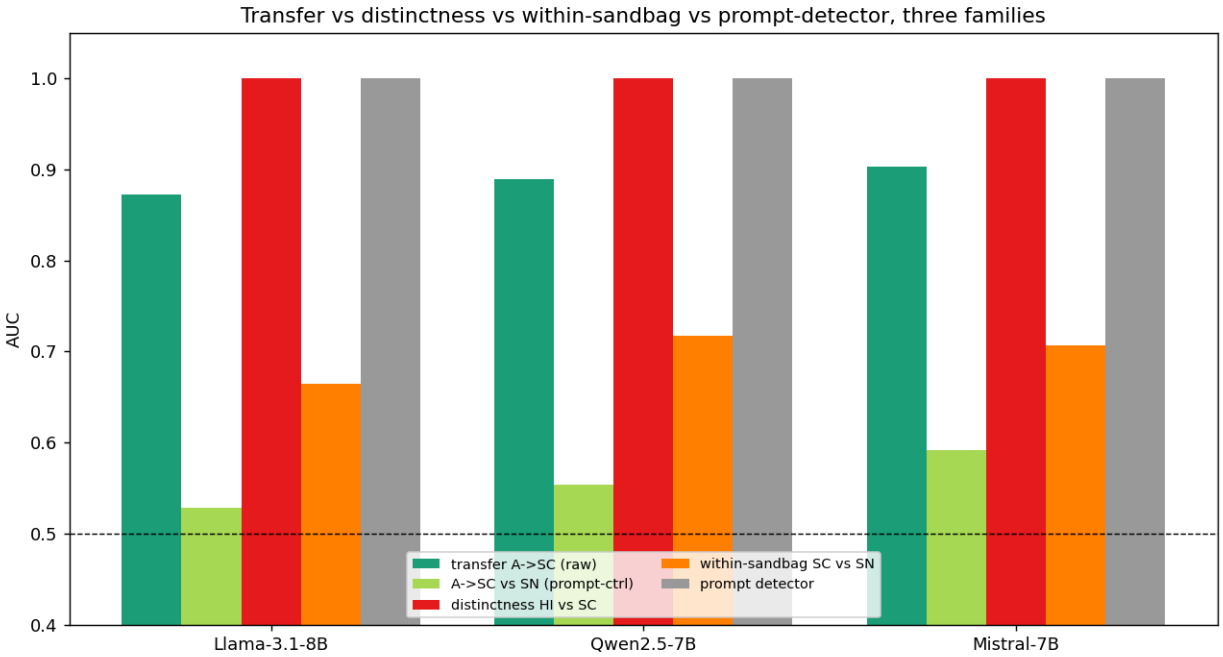

Figure 3: Transfer, distinctness, within-sandbag and prompt AUCs by family

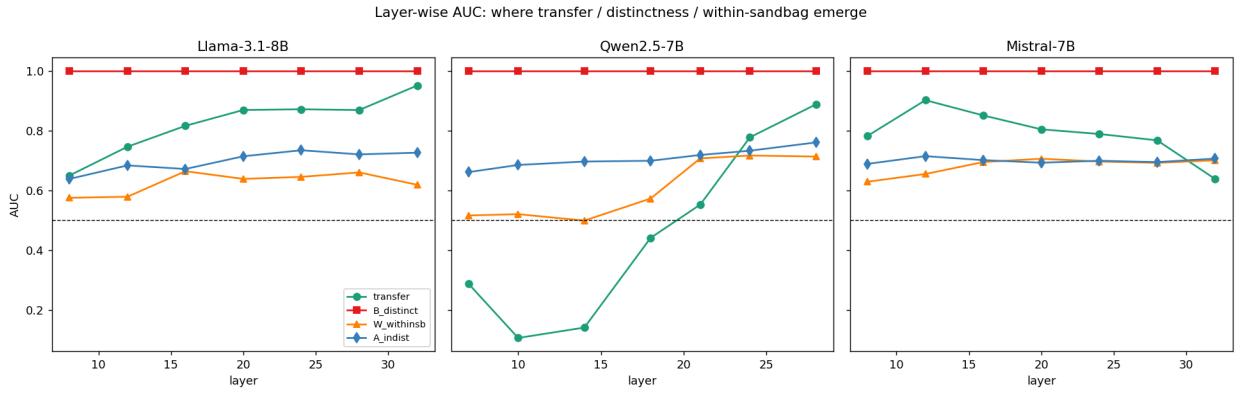

Figure 4: Layer-wise sweep of the probe AUCs

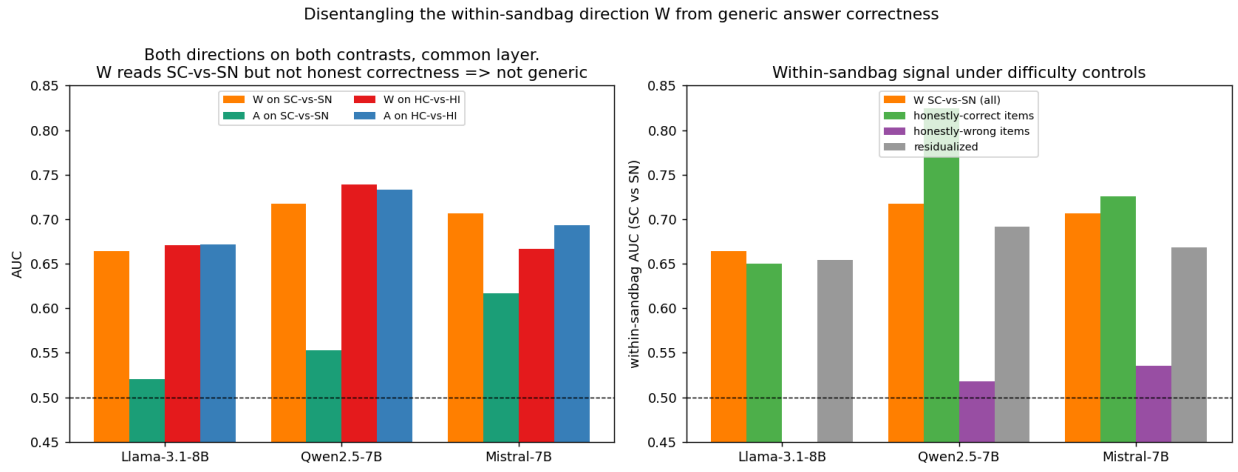

Figure 5: Disentangling the within-sandbag direction from generic answer correctness

this way. A password-locked or fine-tuned sandbagger that removes the prompt while keeping the behavior is needed to measure a prompt-independent signal cleanly.

The scope is narrow. The evidence comes from three small instruct models in the seven to eight billion parameter range, a single multiple-choice benchmark in MMLU, direct single-letter answering with no chain of thought, and one family of last-token linear probes on the residual stream. Whether the findings hold for larger models, generative tasks, reasoning traces, or nonlinear and multi-token readouts is untested.

Finally, even setting the correctness caveat aside, the within-sandbag signal is weak, with AUC near 0.66 to 0.72. That is far from the accuracy a deployable detector would need, so the negative headline does not hinge on the correctness interpretation.