Does a checkable-error probe transfer to confabulation, or is fluent fabrication internally distinct from checkable wrongness?

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 the output does not [@burns2023ccs] [@kadavath2022know]. We ask whether an error-awareness probe trained only on deterministic, objectively scorable mistakes also fires when a model confabulates, inventing a confident answer for an entity that does not exist, or whether fluent fabrication occupies a distinct internal direction. Across three instruct models, Llama 3.1 8B [@grattafiori2024llama3], Qwen2.5 7B [@qwen2024qwen25] and Mistral 7B [@jiang2023mistral], we collect checkable items (arithmetic and capital-city questions split into correct, CC, and genuinely wrong, CI) and confabulation items (fake-entity prompts with no true answer, split into confident fabrication, CF, and abstention, AB, on the identical prompt). A checkable-error probe (CC versus CI) is strongly decodable, AUC 0.8707, 0.9057 and 0.9346, yet it does not transfer to confabulation: scoring CF against AB it reaches only 0.3404 on Qwen, below chance, and 0.5617 on Mistral, at chance [@hendrycks2017baseline]. The reason is that confabulation reads as felt-correct, cleanest on Qwen: there the mean error-score of CF is 0.0041, below even correct answers (CC 0.0910) and far below genuine errors (CI 0.6823), so the model registers its fabrications as more correct than its real correct answers, whereas on Mistral CF sits at 0.3441, above CC (0.1144) but still far below CI (0.7854). A within-fake-entity probe (CF versus AB), which holds the fake-entity prompt constant [@chwang2024androids] [@ji2023survey], separates the two responses near-perfectly, AUC 0.9873 and 0.9780, holding within the book and person entity types that actually confabulate; but a real-control test rescopes what it reads. Genuine confident naming of real entities (RC) projects onto this axis exactly where fabrication does (RC versus AB AUC 0.9920 and 0.9960, CF versus RC only 0.6814 and 0.3296), so the axis decodes a committed specific answer against an abstention and is only partly fabrication-specific. A cross-probe matrix shows a double dissociation, the error probe fails on confabulation and the confabulation probe fails on checkable error (Qwen 0.1334), and the two probe directions are near-orthogonal at every layer (cosine magnitude at most 0.0635). The clean dissociation is strongest in Qwen and only partial in Mistral, where the error probe pushes CF partway toward the error signature (CF versus CC AUC 0.8330) and the confabulation direction carries some error signal (0.7381). The practical implication is that an error-detector built on checkable wrongness is not, on its own, a confabulation detector [@park2024deception] [@pacchiardi2024liar], and the obvious within-fake-entity probe largely tracks answering against abstaining rather than fabrication, so a fabrication-specific signal still has to be isolated [@manakul2023selfcheckgpt] [@kuhn2023semantic].

Setup

For each model we run two families of prompts. Checkable items are arithmetic and capital-city questions that have an objective answer; the model's response is scored automatically into checkable correct (CC) and checkable incorrect (CI), the genuine task errors. Confabulation items name a non-existent country, book, person or place, so there is no true answer; the model either invents a confident answer (confident fabrication, CF) or refuses and says the entity does not exist (abstention, AB). Because CF and AB are responses to the same fake-entity prompts, the CF-versus-AB contrast holds the prompt constant and isolates fabrication from prompt-weirdness and out-of-distribution effects, the key control adopted from the sibling sandbagging study. A real-control group (RC) of genuine entities of the same types checks that the fakes are not simply unanswerable, and serves below as the matched specific-real-answer group for the answer-specificity control of Probe C. Answering is direct, 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 across seven layers per model. All probes are L2 logistic regressions on standardized features with the scaler fit on train folds only, seed 42, scored by ROC-AUC on pooled out-of-fold predictions of a 5-fold stratified split, with 95% confidence intervals from 1000 bootstraps and the layer chosen by cross-validation on training data only [@belinkov2022probing]. Probe A is the checkable-error detector (CC versus CI), Probe C is the within-confabulation detector (CF versus AB), and transfer asks whether Probe A, refit on all checkable

rows, separates CF from AB and from CC.

Probe metrics

model	probe	task	layer	AUC	ci_lo	ci_hi	Cohen d
llama-3.1-8b-instruct	A_CCvsCI	CC_vs_CI	8	0.8707	0.8233	0.9116	1.4612
qwen2.5-7b-instruct	A_CCvsCI	CC_vs_CI	14	0.9057	0.8591	0.9440	2.2168
qwen2.5-7b-instruct	C_CFvsAB	CF_vs_AB	28	0.9873	0.9740	0.9968	4.7292
mistral-7b-instruct-v0.3	A_CCvsCI	CC_vs_CI	24	0.9346	0.9105	0.9566	2.1741
mistral-7b-instruct-v0.3	C_CFvsAB	CF_vs_AB	32	0.9780	0.9581	0.9924	4.2490

A deterministic checkable-error probe is linearly decodable in every model: Probe A reaches AUC 0.8707 on Llama (layer 8, 68 errors against 332 correct, Cohen d 1.46), 0.9057 on Qwen (layer 14, 65 against 335, d 2.22) and 0.9346 on Mistral (layer 24, 139 against 261, d 2.17), squarely in the error-awareness range [@kadavath2022know]. Probe C, the within-confabulation detector, is far stronger still, AUC 0.9873 on Qwen (layer 28, 113 fabrications against 187 abstentions, d 4.73) and 0.9780 on Mistral (layer 32, 100 against 200, d 4.25). The signal for the model is making this up exists and is almost perfectly separable, but, as the next sections show, it lives on its own axis. Llama is omitted from the confabulation probe because it abstains on nearly every fake entity (CF = 1), making it the abstainer baseline rather than a confabulator; the real-control answer rate is 1.0 for all three models, so the fakes are not merely unanswerable.

The checkable-error probe does not transfer to confabulation

model	A on CF vs AB	ci_lo	ci_hi	A on CF vs CC	ci_lo	ci_hi	d(CF,CC)	d(CF,CI)	CF position (CC=0, CI=1)
qwen2.5-7b-instruct	0.3404	0.2849	0.3978	0.2627	0.2182	0.3078	-0.4229	-2.8565	-0.1469
mistral-7b-instruct-v0.3	0.5617	0.4985	0.6246	0.8330	0.7905	0.8732	0.6854	-1.1472	0.3424

Refit on all checkable rows at its best layer and applied to the confabulation rows, Probe A cannot pick fabrications out from abstentions. Its CF-versus-AB AUC is 0.3404 on Qwen, below chance, and 0.5617 on Mistral, at chance. The four-group error-score distribution explains why. On Qwen the mean error-score is 0.0910 for CC, 0.6823 for CI, 0.0041 for CF and 0.0630 for AB: confabulations score essentially zero, lower than the model's own correct answers, so on the error axis a fabrication looks more correct than a genuine correct answer. The placement is captured by the CF position metric, which puts CF at -0.1469 on a scale where CC is 0 and CI is 1, below the correct-answer anchor, and by the large negative effect size against genuine errors, d(CF,CI) = -2.8565. On Mistral the fake-entity prompt elevates both CF and AB on the error axis (means 0.3441 and 0.3819) but to the same level, well below CI (0.7854): the prompt raises the error-score equally for fabricating and for abstaining, so fabrication itself is invisible to the error axis (CF-versus-AB 0.5617). Confabulation does not share the checkable-error signature; fluent fabrication is internally felt-correct, consistently across both models that confabulate, with the caveat that on Mistral the error probe does push CF partway toward the error region relative to correct answers (CF-versus-CC AUC 0.8330, d(CF,CC) = +0.6854), so the felt-correct reading is cleanest on Qwen. The error-score distributions across all four groups are shown in the main figure.

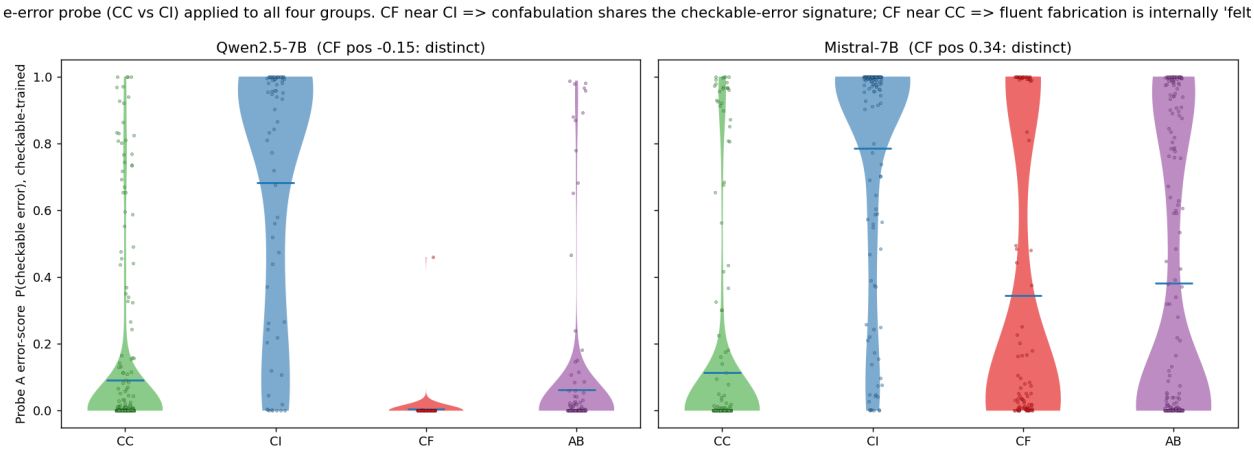


Figure 1: Probe A checkable-error score across the four groups CC, CI, CF and AB, one panel per confabulating model. Genuine errors (CI) sit high; confabulations (CF) sit at or below correct answers (CC), showing fabrication reads as felt-correct on the error axis.

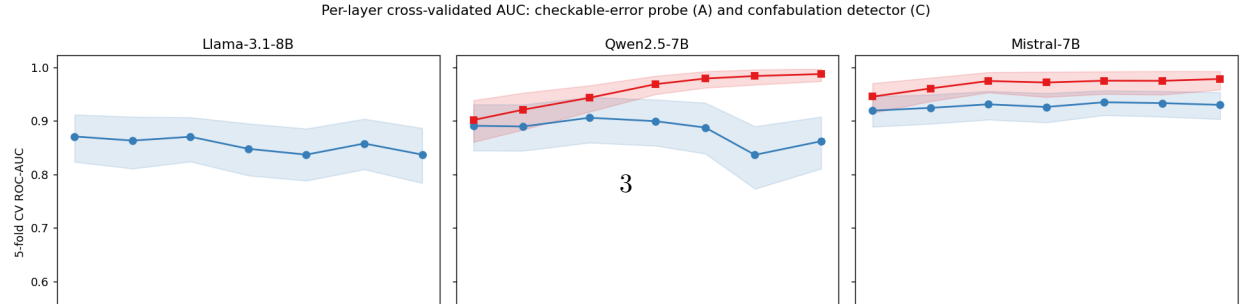
The underlying four-group error-score statistics, the data plotted in the main figure, are given below for each confabulating model, with group sizes, means, medians and 95 percent bootstrap confidence intervals.

Model	Group	n	Mean	Median	CI low	CI high
Qwen2.5 7B	CC	335	0.0910	0.0000	0.0694	0.1177
Qwen2.5 7B	CI	65	0.6823	0.9405	0.5903	0.7711
Qwen2.5 7B	CF	113	0.0041	0.0000	0.0000	0.0123
Qwen2.5 7B	AB	187	0.0630	0.0000	0.0355	0.0959
Mistral 7B	CC	261	0.1144	0.0001	0.0802	0.1504
Mistral 7B	CI	139	0.7854	0.9911	0.7238	0.8433
Mistral 7B	CF	100	0.3441	0.0517	0.2602	0.4311
Mistral 7B	AB	200	0.3819	0.0405	0.3242	0.4410

A double dissociation between the two axes

model	A on CC vs CI	A on CF vs AB	C on CF vs AB	C on CC vs CI
qwen2.5-7b-instruct (L14)	0.9057	0.3404	0.9429	0.1334
mistral-7b-instruct-v0.3 (L24)	0.9346	0.5617	0.9748	0.7381

Evaluated at a common layer, the cross-probe matrix shows a double dissociation. The error probe is strong on its own task and weak off it (Qwen 0.9057 on CC-versus-CI, 0.3404 on CF-versus-AB), and the confabulation probe is strong on its own task and weak off it (Qwen 0.9429 on CF-versus-AB, 0.1334 on CC-versus-CI, far below chance). Each probe fails on the other's contrast, so the two are reading different internal features rather than one shared wrongness signal seen twice. The dissociation is clean on Qwen and only partial on Mistral, where the confabulation direction carries a moderate amount of error signal at this common layer (C on CC-versus-CI 0.7381), hinting at a partial shared sub-component in that model. The per-layer AUC curves for both probes are shown below, and the cross-probe transfer is summarized as a heatmap.



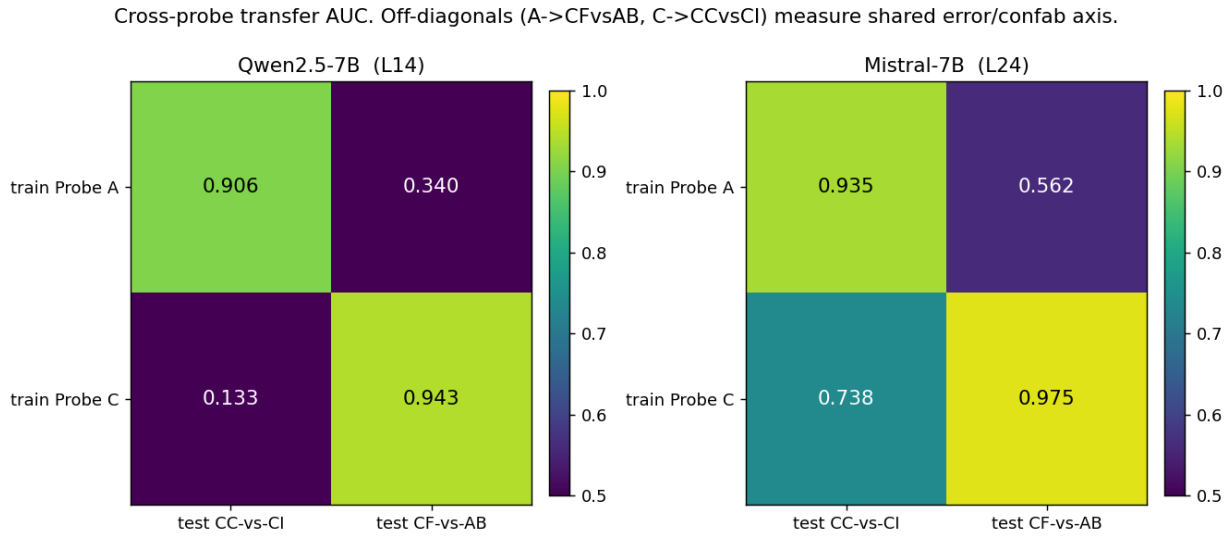


Figure 3: Two-by-two cross-probe transfer AUC per confabulating model: rows are the trained probe (A or C), columns are the test contrast (CC-vs-CI or CF-vs-AB). The strong diagonal and weak off-diagonal are the double dissociation.

model	entity subset	n CF	n AB	AUC book+person	ci_lo	ci_hi	full AUC
qwen2.5-7b-instruct	book+person	100	60	0.9887	0.9705	1.0000	0.9873
mistral-7b-instruct-v0.3	book+person	96	64	0.9797	0.9618	0.9925	0.9780

Because CF and AB share the fake-entity prompt, the near-perfect Probe C separation is not an artifact of prompt-weirdness; it is a robust signal read under a matched prompt control, though the real-control test in the next section shows it tracks committed answering rather than fabrication specifically. The signal is also not an artifact of mixing entity types: restricted to the book and person entities that actually confabulate, Probe C holds at full strength, AUC 0.9887 on Qwen and 0.9797 on Mistral, essentially the full-data values. The fake-country and fake-place items yield too few fabrications per model to test in isolation, so the confabulation result is demonstrated for the entity types that confabulate. Finally, the two probe directions are geometrically distinct. The cosine similarity between the checkable-error weight vector and the confabulation weight vector, both refit on full data at each layer, ranges from -0.0635 to 0.0321 on Qwen and from -0.0393 to 0.0598 on Mistral, near zero at every layer. The checkable-error direction and the confabulation direction are close to orthogonal, not the same wrongness direction read two ways. The cosine-by-layer curve and a two-dimensional projection of the four groups make the geometry visible.

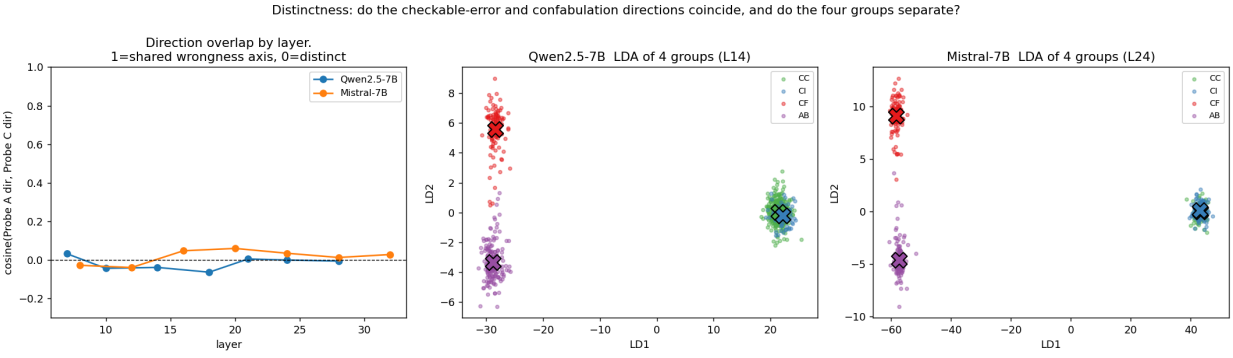


Figure 4: Left, cosine similarity between the checkable-error direction (Probe A) and the confabulation direction (Probe C) at every layer per model, hovering near zero throughout. Right, a two-dimensional discriminant projection of the four groups CC, CI, CF and AB with group centroids, showing the fake-entity prompts and the confabulation split occupy axes distinct from the checkable-error contrast.

A real-control test rescopes the within-fake-entity probe

RC against CF; the robust fact is that RC sits with the committed-answer groups and far from abstention. This does not touch the paper's main result, which rests on the checkable-error probe failing to transfer and on the near-orthogonal directions, but it narrows what the CF-versus-AB probe should be called.

Model	Group	Mean P(CF)	CI low	CI high
Qwen2.5 7B	CF	0.9224	0.8739	0.9608
Qwen2.5 7B	AB	0.0346	0.0146	0.0563
Qwen2.5 7B	RC	0.9819	0.9540	0.9991
Qwen2.5 7B	CC	0.9890	0.9819	0.9946
Mistral 7B	CF	0.9078	0.8560	0.9526
Mistral 7B	AB	0.0443	0.0220	0.0719
Mistral 7B	RC	0.9631	0.8913	0.9998
Mistral 7B	CC	0.9809	0.9686	0.9898

Model	Contrast	AUC	CI low	CI high
Qwen2.5 7B	RC vs AB	0.9920	0.9786	0.9997
Qwen2.5 7B	CF vs RC	0.6814	0.5914	0.7677
Mistral 7B	RC vs AB	0.9960	0.9894	1.0000
Mistral 7B	CF vs RC	0.3296	0.1858	0.4834

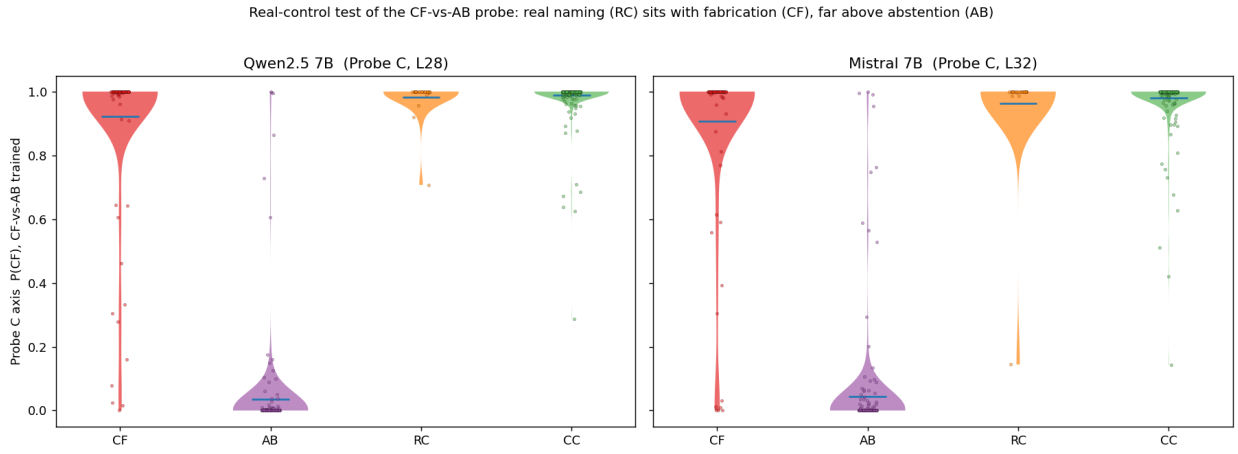


Figure 5: Probe C axis score, P(CF) from the CF-versus-AB probe, across the four groups CF, AB, RC and CC, one panel per confabulating model. Real confident naming (RC) and out-of-domain specific answers (CC) sit with fabrication (CF) far above abstention (AB), showing the axis decodes a committed answer versus an abstention rather than fabrication.

Limitations

The central caveat is that the clean dissociation is strongest in Qwen and only partial in Mistral. On Mistral the checkable-error probe separates CF from CC at AUC 0.8330 with d(CF,CC) = 0.6854, so fabrications drift partway toward the error signature rather than sitting cleanly with correct answers as they do on Qwen, and the confabulation direction carries a moderate error signal at the common layer (C on CC-versus-CI 0.7381). The fully distinct story is therefore a Qwen result that holds only partially on Mistral, and we report both rather than averaging them away. Llama abstains on almost every fake entity (CF = 1), so the confabulation analyses rest on two models. The answer-specificity control rests on 24 real-control items per model, enough to place real naming firmly with the committed-answer groups and far from abstention but

not to resolve its exact ordering against fabrication. Confabulation also concentrates in book and person items; fake-country and fake-place produce too few fabrications per model to test separately, so the result is established for the entity types that confabulate, not for all four. The activations are last-position residual stream at seven layers, read by last-token linear probes, so richer pooling, attention-head-level readouts or nonlinear probes might find structure the layer-level linear probe misses. The evidence comes from three instruct models in the seven-to-eight-billion range under direct single-letter answering, and whether the felt-correct result holds at larger scale, on generative tasks, or with reasoning traces is untested. The probes are read-only, so whether the confabulation direction causally controls fabrication, the natural follow-up by steering along Probe C, is open.

Conclusion

A checkable-error probe trained on deterministic arithmetic and capital-city mistakes does not, by itself, detect confabulation. It is strongly decodable on its own task (AUC 0.87 to 0.93) but at or below chance on fabrication (CF-versus-AB 0.3404 and 0.5617), because confabulation reads as felt-correct, scoring at or below correct answers on the error axis while genuine errors score high. A within-fake-entity probe, with the fake-entity prompt held constant, separates the committed fabrication from abstention near-perfectly (0.9873 and 0.9780) and holds within entity type, but a real-control test shows that this axis largely tracks committed answering versus abstaining rather than fabrication itself: genuine confident naming of real entities scores with fabrication, not with abstention (CF-versus-RC AUC 0.6814 and 0.3296, against RC-versus-AB 0.9920 and 0.9960). The cross-probe matrix shows a double dissociation between the error axis and this commitment axis, and the two directions are near-orthogonal at every layer. Fluent fabrication is internally distinct from checkable wrongness, cleanly in Qwen and partially in Mistral, which means that detecting fabrication for oversight requires a fabrication-specific signal rather than a repurposed error detector, and that signal is not delivered by the obvious within-fake-entity probe either [@park2024deception] [@kuhn2023semantic] [@xiong2023canllms] [@tian2023justask] [@lin2022truthfulqa] [@zou2023repe] [@li2023iti].