PARALLAX: AUTOMATIC LAYER SELECTION FOR INTERNAL-STATE HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models hallucinate with confidence, producing text that is fluent, authoritative, and wrong. In medicine and other high-stakes uses, that failure can cause direct harm. Detectors that read a model's own hidden activations promise to catch it before the text reaches a user, and they report high AUROC on standard benchmarks. Those performance numbers are fragile for two reasons. First, in three of the four corpora we study, the detector input already contains the candidate answer being judged, and a sentence-embedding classifier with no access to the model reaches 0.65 to 0.79 AUROC on them. Second, existing detectors are static because they read from a layer selected in advance, often the final layer. That layer is rarely the right one. Cross-validation inside the training data did not select it in any of our 1,000 training folds. The depth that works best also moves from one model to the next, although within a model it barely moves across corpora. **PARALLAX** searches every layer under cross-validation that never touches the test data, and selects the ones that score best. The effect is large. Across five open-weight models and four corpora, one selected layer beats last-layer probing on every one of those twenty model and corpus pairings, by as much as 0.117 AUROC. It carries to full scale. On the complete corpora, up to 15,090 examples and thirty times the data, it holds on all 100 held-out splits and the largest gain reaches 0.123 AUROC. PARALLAX has the highest AUROC of any detector that reads the model only once, on thirteen of the twenty tests. It keeps its advantage in 58 of 60 cross-corpus transfers, where it was tested on corpora it was not trained on. At a hundred labelled examples and below, it beats the strongest detector that scores tokens individually on every seed. This shows that choosing the layer, with no change to the classifier and no extra forward pass, is enough to put a plain linear probe ahead of every published detector that reads the model once. Which layer a detector reads is therefore a design decision worth making, not a constant fixed in advance.

1 INTRODUCTION

Large language models are now used to answer medical, legal, and scientific questions, and their errors are hard to spot because they read like correct answers (Zhang et al., 2023; Huang et al., 2025; Ji et al., 2023). Retrieval grounds a model in supplied text (Lewis et al., 2020) without removing the failure (Sun et al., 2024; Zhang et al., 2024). The model's output probabilities are a poor guide to this, because a model can assign high probability to a claim it has no support for. Its internal activations carry information the output distribution does not expose, including whether the fact being asserted is actually stored in the model (Petroni et al., 2019; Meng et al., 2022; Mallen et al., 2023). A common response is therefore to read those activations directly. A light classifier takes the states the model produced for an answer and predicts whether that answer is wrong (Azaria & Mitchell, 2023; Chen et al., 2024; Su et al., 2024; Du et al., 2024; Niu et al., 2025). These detectors need no extra generations and can run next to the model at deployment time.

Two problems stand between these detectors and their reliable use. The first is how they are evaluated. In three of the four corpora we study, the detector input already contains the candidate answer being judged, so a classifier that only reads input text, with no access to the model, reaches AUROC 0.65 on TruthfulQA, 0.68 on MedHallu and 0.79 on HaluBench (Section 5). A high score on such a corpus says little about the model's internal states unless it exceeds this text-only control by a significant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

margin. Features must also be computed on exactly the text the label describes, the stored response or the candidate answer. If the probed model is asked to generate fresh text and the features are read from that instead, the label and the features no longer refer to the same words, and a detector can appear accurate without ever having read the text it is being judged on. The second problem is that existing internal-state detectors are static. Each reads one fixed layer, or a fixed set of layers, chosen once and reused for every model and corpus. SAPLMA (Azaria & Mitchell, 2023) compares a few preset depths, and HalluShift (Dasgupta et al., 2025) compares two fixed layers. We find that the most informative layers sit in the middle of the network, at a depth that changes from model to model but stays put across corpora within a model: layer 32 or 33 of Llama-3.3-70B, and layer 43 or 45 of DeepSeek-R1-32B, are selected on all four corpora (Section 6.3). The choice is therefore a property of the model, so any preset depth is wrong on some models, and the search only has to be run once per model. Nothing is given up to make it. Every supervised detector of this kind, SAPLMA, PRISM and HaMI included, already fits a classifier on labelled data from the target corpus, and PARALLAX spends those same labels on the layer as well, so it asks for no data a fixed-layer probe does not already need. The search itself costs minutes of CPU on states that are extracted once, inference is unchanged at one forward pass and a linear classifier, and where a corpus has no labels at all a choice made on another corpus still beats the fixed layer (Section 7).

We introduce **PARALLAX**, a probe that is not tied to any layer (Figure 1).[1] It selects the layers to read by greedy forward selection, scoring every candidate by cross-validated AUROC on the training data only. The selection is run separately for each model and corpus. That is the conservative choice, since the corpus turns out to change the answer very little. With one selected layer, it is a linear probe on that layer; with $k > 1$ layers it classifies the differences between their representations. We report $k = 1 \ldots 4$ separately, because the contribution is which layers to read.

We evaluate under a corrected protocol on five open-weight models from four families (8B to 70B parameters) and four corpora, against seventeen published detectors that we rerun on exactly the labeled text, with nested cross-validation so that *no choice ever sees test data*. We make two contributions, and each needs the other. The first is a corrected re-evaluation of published detectors, which shows how much of their reported accuracy the benchmarks hand them for free, and which corpora are still able to separate one method from another. The second is PARALLAX, which uses those corpora to show that the layer a probe reads decides whether it works.

- **A corrected re-evaluation of published detectors:** Scored on the text their labels describe, sampling- and uncertainty-based detectors (semantic entropy, SEPs, MIND, INSIDE, ACT) stay between 0.27 and 0.61 AUROC, label-free internal-state scores often fall below chance, and supervised probes work but depend on a layer fixed in advance. The same pass measures how much of each corpus an input text-only classifier can already predict. This separates RAGTruth and TruthfulQA, which still tell methods apart, from HaluBench and MedHallu, which are saturated (Section 6).
- **PARALLAX**: automatic layer selection. Selecting a layer beats reading a fixed one in all 20 model–corpus settings and on all 200 held-out seeds, by as much as 0.117 AUROC. The bootstrap intervals over examples exclude zero everywhere. It carries to full scale, holding in all 20 settings and on all 100 held-out splits of the complete corpora, up to 15,090 examples and thirty times the data, where the largest gain reaches 0.123 AUROC. It also beats every published method that tunes its own depth. On RAGTruth and TruthfulQA it is ahead of SAPLMA's five preset depths in all ten of those model–corpus settings, and it beats a single relative depth tuned once across the whole benchmark in 17 of the 20. The choice transfers across corpora. Selected on one corpus and applied to another, it still beats the fixed-layer probe in 58 of 60 transfers, and within a model the same band of layers is chosen on all four corpora, so the search runs once per model. It is also the cheapest tuning on offer, consuming no labels beyond those every supervised detector already trains on, costing minutes of CPU on activations extracted once, and leaving inference at one forward pass and a linear classifier (Section 7).

2 Related Work

Internal-state detectors: SAPLMA (Azaria & Mitchell, 2023) trains a feed-forward classifier on the hidden state of one layer; its authors compare five fixed layers of 32-layer models and report

[1] Code: https://anonymous.4open.science/r/llm-fuzzing-7613

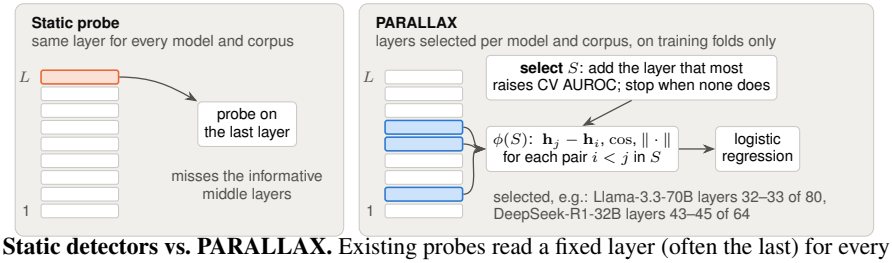

Figure 1: **Static detectors vs. PARALLAX.** Existing probes read a fixed layer (often the last) for every model and corpus. PARALLAX scores every layer on training folds, keeps the best one, and adds further layers only while cross-validated AUROC improves. With $k > 1$ layers it classifies the differences between their pooled representations.

that a deep one is usually best. INSIDE (Chen et al., 2024) scores the covariance of states across sampled answers, MIND (Su et al., 2024) trains on pseudo-labels, and HaloScope (Du et al., 2024) scores the projection of a state onto a subspace learned from unlabeled data. SEPs (Kossen et al., 2024) regress semantic entropy from a hidden state, PRISM (Zhang et al., 2025a) adds a truthfulness prompt before reading the state, HalluShift (Dasgupta et al., 2025) measures the shift between two layers, HaMI (Niu et al., 2025) applies multiple-instance learning to token-level states, and IRIS (Srey et al., 2025) trains on the embedding of a self-verification prompt. Detectors built on other internal signals make the same commitment: ReDeEP (Sun et al., 2024) contrasts parametric and retrieved knowledge at chosen layers, the ICR probe (Zhang et al., 2025b) tracks hidden-state dynamics across a fixed span, and entity-level detectors read a fixed depth (Obeso et al., 2025; Ferrando et al., 2025). In all of these methods the layer is a hyperparameter set in advance, not something the method selects for each model (Preiß, 2025). Orgad et al. (2025) shows that truthfulness signals are concentrated in particular tokens and do not transfer well across datasets.

Which layers carry the signal: A longer line of work reads concepts directly from activations, whether to locate factual associations (Meng et al., 2022), to recover latent knowledge without supervision (Burns et al., 2023), or to steer generation along a direction found in activation space (Zou et al., 2023; Li et al., 2023b; Turner et al., 2023; Panickssery et al., 2023). The probing work finds that the truth-related directions are linearly readable in intermediate layers (Marks & Tegmark, 2024; Belinkov, 2022), and that intermediate layers often give better representations than the final for many downstream tasks (Skean et al., 2025; Tenney et al., 2019; Wang et al., 2025b). Zhang et al. (2026) reports that the layers that matter most for reinforcement-learning fine-tuning concentrate in the middle of the stack. Our results agree with this observation for hallucination detection, and add that the best depth is a property of the model rather than the corpus, so no fixed depth serves every model (Section 6.3).

Sampling-based detectors: SelfCheckGPT (Manakul et al., 2023) checks an answer against the model's own samples, semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) measures disagreement among sampled answers, and P(True) (Kadavath et al., 2022) asks the model to judge an answer. These need several extra generations per query.

Evaluation of hallucination detectors: Janiak et al. (2025) show that reported progress depends on the metric used to label generations (ROUGE vs. human judgment). Benchmark construction has drawn similar scrutiny: summarisation-consistency benchmarks disagree with each other and with human labels (Laban et al., 2022; Maynez et al., 2020), and generated-answer benchmarks pair a question with a prepared correct or corrupted answer by construction (Li et al., 2023a), which is the pattern we quantify here. Finer-grained protocols score individual facts instead of whole answers (Min et al., 2023; Quevedo et al., 2024). We study a complementary issue, what the detector input contains and which text its features describe.

3 PARALLAX: Automatic Layer Selection

PARALLAX makes the layer a decision taken per model and corpus rather than a constant fixed by the designer. It has three parts. A feature map that turns a set of layers into a vector, a greedy search picks that set using training data alone, and a logistic probe is fit on the result. Only the search differs from a conventional fixed-layer probe. The feature map and the classifier are deliberately plain, so that whatever the method gains can be attributed to where it reads rather than to how it reads.

Representations: We index the decoder layers $1, \dots, L$ and write $\mathbf{h}_\ell \in \mathbb{R}^d$ for the output of layer ℓ averaged over the tokens of the scored text. For a pair of layers, we define how the representation changes between them,

$$\psi(\ell_i, \ell_j) \;=\; \left[\; \mathbf{h}_{\ell_j} - \mathbf{h}_{\ell_i} \;,\;\; \cos(\mathbf{h}_{\ell_i}, \mathbf{h}_{\ell_j}) \;,\;\; \|\mathbf{h}_{\ell_j} - \mathbf{h}_{\ell_i}\|_2 \;\right] \;\in\; \mathbb{R}^{d+2}. \tag{1}$$

A selection is a set of layers $S = \{\ell_1 < \cdots < \ell_k\}$, and the probe reads

$$\phi(S) \;=\; \begin{cases} \mathbf{h}_{\ell_1} \in \mathbb{R}^d & k = 1, \\ \left[\, \psi(\ell_i, \ell_j) \,\right]_{i<j} \in \mathbb{R}^{\binom{k}{2}(d+2)} & k > 1, \end{cases} \tag{2}$$

the pair features of all $\binom{k}{2}$ pairs concatenated. Features are standardized and passed to an L2-regularized logistic regression (C=1, balanced class weights). Algorithm 1 selects S on the training data D. It scores every single layer by inner cross-validated AUROC (referred as CVAUROC in Algorithm 1) and keeps the best. It then tries adding each remaining layer, keeps the addition with the highest inner AUROC, and stops when no addition improves it or when $|S| = k_{\max}$. The procedure places no constraint on which layers are chosen or how many ($1 \le k_{\max} \le L$), and Section 6.4 reports it at exactly $k = 1 \dots 4$. The mechanism is identical at every k; what changes with the corpus is how much a second layer adds. We carry both $k = 1$ and the adaptive rule with $k_{\max} = 4$, which chooses k without being told, through the results.

1: $a_\ell \leftarrow \mathrm{CVAUROC}(\phi(\{\ell\}), D)$ for every layer $\ell = 1, \dots, L$
2: $S \leftarrow \{\arg\max_\ell a_\ell\}$, $\quad a^* \leftarrow \max_\ell a_\ell$
3: **while** $|S| < k_{\max}$ **do**
4: $\quad \ell' \leftarrow \arg\max_{\ell \notin S} \mathrm{CVAUROC}(\phi(S \cup \{\ell\}), D)$, $\quad a' \leftarrow$ its score
5: $\quad$ **if** $a' \le a^*$ **then break** (omitted in the fixed-k variant)
6: $\quad S \leftarrow S \cup \{\ell'\}$, $\quad a^* \leftarrow a'$
7: **return** S and a logistic probe fit on $\phi(S)$ over all of D

Algorithm 1: PARALLAX layer selection on training data D (inner 5-fold CV).

Selection needs $O(k_{\max} L)$ probe fits per inner fold, on pooled states that are extracted once. At test time PARALLAX needs one forward pass, the pooled states of at most $k_{\max}$ layers, and a linear classifier. It needs no sampled generations.

4 Experimental Setup

Models: Llama-3.3-70B-Instruct and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), Qwen3-8B (Yang et al., 2025), and Phi-4 (Abdin et al., 2024) (80, 64, 32, 36, and 40 layers). The two largest run in 4-bit NF4 quantization on one GPU, the others in bfloat16.

Single-pass baselines: (all five models). SAPLMA (Azaria & Mitchell, 2023) in two forms, a logistic probe on the last layer, and with its five candidate depths (the last layer, the layers 4, 8 and 12 before it, and the middle layer, scaled to each model), the layer chosen by inner cross-validation, with either a logistic or the original 256-128-64 feed-forward classifier. HaloScope (Du et al., 2024), HalluShift (Dasgupta et al., 2025), PRISM (Zhang et al., 2025a), and HaMI (Niu et al., 2025), the layer contrast used by DoLa (Chuang et al., 2024), mean token log-probability, P(True) (Kadavath et al., 2022), and a self-verification score (P("No") normalized over "Yes" and "No" after asking the model whether the answer is correct, in the spirit of IRIS (Srey et al., 2025)). Label-free scores (†) keep the sign defined by their method. Rows marked ‡ are simplified versions that keep the core feature of the method. Appendix A.2 gives the details.

Sampling-based baselines: (Llama-3.3-70B and DeepSeek-R1-32B). These methods are defined over a model's own sampled generations, but these corpora supply the text to be judged, so scoring them at all requires sampling from the prompt and comparing those samples against the supplied text. That is a property of the task these corpora pose rather than of the methods, and their numbers should be read accordingly. We draw N=10 samples per question from the prompt without the candidate answer and score the labeled text with SelfCheckGPT-NLI (Manakul et al., 2023), semantic entropy (Farquhar et al., 2024), SEPs (Kossen et al., 2024), MIND (Su et al., 2024), ACT (Wang et al., 2025a), INSIDE (Chen et al., 2024), and HaMI with its official code, which weights token states by the sampled answers. We also run the official HaMI without these weights.

Table 1: Corpora. For each corpus we use a fixed random sample of 500 examples. The detector input is the prompt followed by the scored text; the label describes the scored text.

Corpus	Task	Prompt	Scored text	Halluc./total
RAGTruth (Niu et al., 2024)	RAG summaries and QA	source + instruction	stored LLM response	153 / 500
TruthfulQA (Lin et al., 2022)	factual QA	question	candidate answer	246 / 500
HaluBench (Ravi et al., 2024)	multi-domain QA	passage + question	candidate answer	250 / 500
MedHallu (Pandit et al., 2025)	medical QA	context + question	candidate answer	250 / 500

Corpora: Table 1 lists the four corpora. RAGTruth provides responses written by several LLMs for retrieval-augmented tasks, with expert labels. It is the one corpus whose labels were not built by pairing a question with a prepared answer, and the one whose text the probed model did not write, both consequences of expert token-level annotation being expensive enough that responses come from a fixed pool of generators. Every detector here is compared in that same setting. In TruthfulQA, HaluBench, and MedHallu, each example pairs a question (and, for HaluBench and MedHallu, a supporting passage) with a candidate answer that is either correct or hallucinated. Appendix A.1 shows the exact formats.

5 Evaluation Protocol

Every label describes a specific text, RAGTruth's stored response or the candidate answer. All hidden-state methods therefore run one forward pass over the prompt followed by that text and pool states over the text's tokens only. Prompt and text are tokenized separately and concatenated, so the pooled positions match the labeled tokens exactly. Sampling-based methods still draw samples from the model, but score the labeled text against them (Section 4).

Protocol: Figure 3 in Appendix A.3 shows the evaluation. The outer loop is stratified 5-fold cross-validation (80% train, 20% test per fold), repeated with 10 seeds. Every choice a method makes (PARALLAX's layers, SAPLMA's layer, a baseline's checkpoint) is made inside the outer training set, using an inner stratified 5-fold cross-validation (64% inner train, 16% inner validation of all data). The chosen configuration is then refit on the full outer training set and scored once on the held-out 20%. We pool the test-fold predictions into one AUROC per seed and report the mean and standard deviation over seeds. All methods use the same outer splits, so we also report paired seed wins.

6 Results

Because the candidate answer is part of the input, part of the label can be predicted from its wording alone. We measure this with TxTemb: a logistic regression on a frozen sentence embedding (all-MiniLM-L6-v2) of the scored text, trained with the same labels and splits as every probe. It has no access to the model. TxTemb reaches 0.65 AUROC on TruthfulQA, 0.68 on MedHallu, and 0.79 on HaluBench, and 0.52 on RAGTruth (Table 2). A detector that stays near these values gives no evidence of reading anything beyond the text. The two corpora where this control is strongest, HaluBench and MedHallu, are also the two where every supervised probe sits between 0.94 and 0.995, so they no longer separate methods; RAGTruth and TruthfulQA carry most of the discriminating power in what follows.

6.1 Existing detectors

Table 2 compares single-pass detectors. The decoding and prompting scores are weak. The DoLa layer contrast is near chance everywhere, log-probability is above chance only on RAGTruth (0.62) and inverted on MedHallu (0.34), and self-judgement reaches 0.64–0.93 but stays below the supervised probes. HaloScope is weakest, 0.52 to 0.76, which is expected of a method designed to work without labels even though we give it 100 labelled validation examples per fold. HalluShift reaches 0.63 to 0.91 and PRISM 0.62 to 0.98. The pattern among the supervised probes is itself informative. SAPLMA rises from 0.635 to 0.686 on RAGTruth, and from 0.835 to 0.918 on TruthfulQA, when it is allowed to choose among its five candidate depths instead of reading the last layer, while swapping its logistic classifier for the original 256-128-64 network changes little. PARALLAX is at least as good as every detector in the table on all four corpora, and has the highest AUROC in 15 of the 20 model–corpus settings; counting the token-level reader of Appendix E.1, which is not a pooled-state

Table 2: **Single-pass detectors**, AUROC averaged over the five models (per-model values in Appendix C.1). † label-free, sign as defined by the method. HaloScope, HalluShift and PRISM are run from their released code, and Appendix D.2 compares these official runs with simplified versions of the same methods. Last row: number of models on which PARALLAX has the highest AUROC among all detectors in the table.

Method	RAGTruth	TruthfulQA	HaluBench	MedHallu
Text only				
TxTemb	0.523	0.647	0.789	0.676
Published single-pass detectors, official code				
SAPLMA, last layer	0.635	0.835	0.915	0.973
SAPLMA, 5 depths (LR)	0.686	0.918	0.948	0.993
SAPLMA, 5 depths (MLP)	0.649	0.911	0.951	0.990
HaloScope	0.517	0.559	0.694	0.759
HalluShift	0.631	0.682	0.868	0.911
PRISM	0.623	0.884	0.937	0.984
Prompting and decoding scores				
Layer contrast (DoLa)†	0.522	0.488	0.554	0.506
LogProb†	0.624	0.560	0.470	0.343
P(True)†	0.593	0.696	0.766	0.782
Self-verification†	0.635	0.766	0.879	0.934
Ours				
PARALLAX	0.712	0.927	0.952	0.993
PARALLAX is best (models)	5/5	5/5	3/5	2/5

detector, it leads in 13. When the corpora provide sufficient headroom to distinguish among methods, the advantage is decisive. On RAGTruth and TruthfulQA, PARALLAX outperforms the strongest baseline in all ten model and corpus combinations, achieving average AUROC gains of 0.026 and 0.009, respectively. On the saturated corpora the five settings where another probe edges ahead are decided by 0.0001 to 0.003, with bootstrap intervals containing zero in every one (Appendix B.1). There the supervised detectors are equivalent, and PARALLAX reaches that ceiling without being told where to read. The central comparison examined in this paper, however, never results in a tie. A layer can be fixed in advance in three ways, and PARALLAX beats all three. Against the last layer, which SAPLMA reads, it wins in all 20 settings and on all 200 held-out seeds, with mean gains of $+0.012$ to $+0.118$ AUROC (Appendix C.2). Against a fixed set of four taps at 60, 70, 80 and 85% depth it wins all 20. Against a single relative depth tuned once, it wins 17 of the 20 (Section 6.5). Resampling examples rather than splits leaves the gain over the last layer with a 95% interval above zero in all 20 settings (Appendix B.1).

Sampling-based detectors: Table 25 in Appendix E.2 reports detectors that need sampled generations. Semantic entropy, SEPs, MIND and INSIDE are at chance on all four corpora, and ACT falls well below it, because they measure how uncertain the model is about the question rather than whether a given candidate is wrong. SelfCheckGPT, which checks the labeled text against the samples, reaches 0.72–0.76. HaMI is the strongest method we evaluate, ahead of PARALLAX on RAGTruth ($+0.013$ and $+0.019$ on the two largest models) and behind it on the other three. It reads the same depth PARALLAX selects but scores tokens individually (Appendix E.1).

6.2 Full search and preset depths

SAPLMA with its five candidate depths is the only published baseline that also chooses where to read, so it isolates what a full search buys over a preset grid. Table 3 pairs the two methods seed by seed. PARALLAX is ahead in 15 of the 20 settings and 144 of the 200 seeds, and the split across corpora is sharp. On RAGTruth and TruthfulQA, where the supervised probes are far from ceiling, PARALLAX wins all ten settings by $+0.002$ to $+0.071$ AUROC, and in seven of those ten settings it is ahead on every one of the 10 seeds. On the saturated corpora the two are close. Eight of those ten differences are under 0.002 either way, the largest is -0.0025, and depending on the setting PARALLAX is ahead on anywhere from 1 to 9 of the 10 seeds, with no consistent winner. Five preset depths are a serviceable heuristic on a saturated corpus, and they leave AUROC unread when the task is hard. What the full search buys is placement rather than breadth. Searching ten depths spread from quarter-depth to the output recovers the gain almost exactly, and twenty is indistinguishable

Table 3: **A full search against five preset depths.** PARALLAX minus the best of SAPLMA's two five-depth variants, per setting, with PARALLAX's paired seed wins.

	RAGTruth		TruthfulQA		HaluBench		MedHallu	
Model	Δ	wins	Δ	wins	Δ	wins	Δ	wins
Llama-3.3-70B	+0.071	10/10	+0.020	10/10	+0.005	10/10	+0.001	9/10
DeepSeek-R1-32B	+0.019	10/10	+0.005	10/10	+0.002	6/10	−0.000	5/10
Llama-3.1-8B	+0.013	10/10	+0.002	8/10	−0.003	4/10	+0.001	8/10
Qwen3-8B	+0.026	10/10	+0.005	8/10	−0.001	5/10	−0.001	1/10
Phi-4	+0.002	6/10	+0.003	8/10	+0.001	5/10	−0.001	1/10

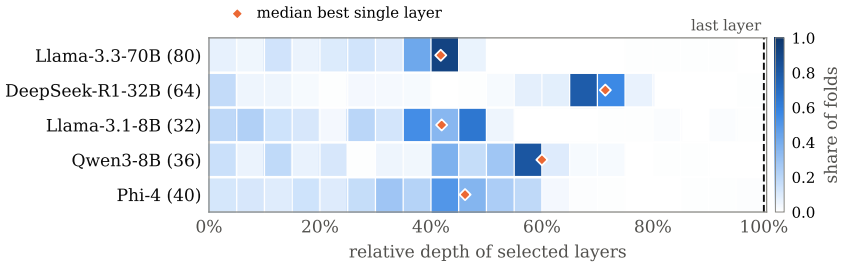

Figure 2: **Selected layers by relative depth.** For each model, the share of held-out folds (50 per corpus, 4 corpora) in which a layer is part of the selected set. Dashed line: last layer.

from searching every layer (Appendix D.1); SAPLMA's five sit in the deep half and miss the band the selection prefers.

6.3 Selected layers

Figure 2 shows where the selected layers lie. The last layer is never the best single layer in any of the 1,000 folds. Selections concentrate in the middle of the network, at a depth that depends on the model. The best single layer sits at about 40% of the depth for Llama-3.3-70B, Llama-3.1-8B, and Phi-4, about 60% for Qwen3-8B, and about 70% for DeepSeek-R1-32B. Within a model the choice is stable across corpora (for example, layer 32 or 33 of Llama-3.3-70B and layer 43 or 45 of DeepSeek-R1-32B on all four corpora). When PARALLAX adds layers, it often pairs the middle layer with a shallow one (0–20% depth), and rarely with a layer in the last third of the network. This agrees with the original SAPLMA finding that an intermediate layer works best on 32-layer models (Azaria & Mitchell, 2023), and with evidence that intermediate layers carry the most useful representations for other tasks (Skean et al., 2025; Zhang et al., 2026). A single fixed set of depths cannot match every model, and SAPLMA's five candidate depths all lie at or beyond the middle of the network, deeper than Llama-3.3-70B's best layers (29–34 of 80).

6.4 Number of layers

Table 4, and Figure 4 in Appendix C.3, force the selection to use exactly k layers, with no early stopping, so $k = 1$ is the best single layer and each further step adds the layer that most raises inner-CV AUROC. Two things follow. First, every k beats the last-layer probe in all 20 settings and on all 200 held-out seeds, and $k = 1$ alone gains between +0.013 and +0.117 AUROC. Second, how much the extra layers add depends on the corpus. On RAGTruth, whose labeled text is a long free-form response, AUROC rises from 0.700 at $k = 1$ to 0.715 at $k = 4$, and the adaptive rule beats one layer on all five models by +0.008 to +0.020 with 9 or 10 of 10 seeds. On the three corpora with short candidate answers the four values sit within 0.002 of each other, and across all 20 settings one selected layer and the adaptive rule are indistinguishable, 104 of 200 seeds with a median difference of −0.0001. Both use the same selection, so this concerns the features rather than the layer choice. The pairwise differences help when the scored text is long and cost nothing when it is short. Choosing k per fold on training data, and tuning the regularization strength, both change nothing (Appendix C.3). What the probe reads matters more than how it is regularized or how many layers it stacks.

Table 4: **Exactly k layers.** AUROC averaged over the five models when the selection is forced to k layers (no early stopping). Per-model values with min–max over seeds in Appendix C.3.

Layers	RAGTruth	TruthfulQA	HaluBench	MedHallu
$k = 1$	0.700	0.928	0.952	0.993
$k = 2$	0.707	0.926	0.952	0.993
$k = 3$	0.713	0.927	0.951	0.992
$k = 4$	0.715	0.927	0.952	0.991
Adaptive k (PARALLAX)	0.712	0.927	0.952	0.993
SAPLMA, last layer	0.635	0.835	0.915	0.973

6.5 A SINGLE GLOBAL DEPTH

Section 6.3 shows that the corpus barely moves the selected layer within a model, which invites the obvious follow-up. If the corpus does not move it, could one relative depth, fixed once and applied to every model, replace the search altogether? It cannot, because the depth still moves across models. We test this directly. For each of 15 relative depths from 20% to 90% we score a probe on that single layer under the same splits, then choose the depth leave-one-setting-out, taking the one with the best mean AUROC over the other 19 settings so that no setting picks its own. PARALLAX beats that globally tuned depth in 17 of the 20 settings, by 0.028 on average and up to 0.136, and the three exceptions differ by at most 0.002. The best depth per setting moves between 40% and 75%, and that movement is between models rather than between corpora. Within a model it shifts by at most one step of the grid on four of the five models, and on the fifth the depth curve is flat enough that the argmax carries little information. A depth tuned over the pool has to serve stacks whose best depths lie 35 points of relative depth apart, and it suits none of them. On DeepSeek-R1-32B RAGTruth the leave-one-out depth reaches only 0.602, below even the last-layer probe. Using every layer is also worse than selecting. A probe on all layers concatenated loses in 11 of the 12 settings where its width is tractable, by 0.014 on average.

6.6 LABEL EFFICIENCY

Token-level annotation is the expensive part of building one of these detectors, so a method is judged by two numbers, how well it scores with 400 labelled examples and how fast it degrades with fewer. We subsample each training fold to 50, 100 and 200 examples, keeping the same outer splits and re-running PARALLAX's layer selection inside each budget, and compare against HaMI, the strongest detector on RAGTruth. The two methods cross over at the same place on both models. With 50 labels PARALLAX reaches 0.595 and 0.597 while HaMI is near chance at 0.547 and 0.537; with 100 labels the gap is $+0.071$ and $+0.066$; at both budgets PARALLAX wins all 10 seeds on both models. HaMI catches up at 200 and passes at 400. The advantage comes from the layer and not from using a linear model: the same probe read from the last layer never beats HaMI at any budget, reaching 0.531, 0.568, 0.599 and 0.658 on Llama-3.3-70B against HaMI's 0.547, 0.589, 0.668 and 0.758. A logistic probe on a few selected layers fits d parameters by convex optimisation, while a multiple-instance network trained for 100 epochs over token bags needs several hundred labelled examples before its extra capacity pays for itself. In the regime where annotation is scarce, which is the regime a new corpus starts in, selecting layers is worth more than reading tokens.

6.7 FULL CORPUS SIZE

The main experiment uses 500 examples per corpus, the size at which token-level annotation is usually available, which leaves open whether selection is just exploiting a small sample. We repeat it on the complete corpora, all of TruthfulQA (1,634), MedHallu (2,000), HaluBench (14,340) and RAGTruth (15,090), the last two thirty times the size used above. Fitting a probe costs more at this scale, so each seed takes one stratified 80/20 split and chooses the layer by inner cross-validation on the 80% alone, five seeds per model.

Selecting the layer beats reading the last one in all 20 settings and on all 100 held-out splits, by $+0.004$ to $+0.123$ AUROC (Table 5). The comparison with SAPLMA's five preset depths moves with corpus size. At 500 examples a full search was ahead in 15 of 20 settings and by as much as $+0.071$; at full size it leads in 17 of 20, but ten of those differences fall below 0.001, the median is $+0.0006$ and the splits go 72/100. TruthfulQA is the exception, holding $+0.007$ to $+0.020$. More labelled data narrows the gap between searching every layer and searching five well-placed ones,

Table 5: **Full corpus size.** AUROC for PARALLAX (k=1), the last-layer probe (last) and SAPLMA's five depths (S-5), five seeds, all five models; RAGTruth and HaluBench are 30× Table 2. Per-cell deviations in Appendix C.4.

	RAGTruth (15,090)			TruthfulQA (1,634)			HaluBench (14,340)			MedHallu (2,000)		
Model	ours	last	S-5	ours	last	S-5	ours	last	S-5	ours	last	S-5
Llama-3.3-70B	0.8559	0.8361	0.8517	0.9524	0.8881	0.9327	0.9880	0.9820	0.9869	0.9969	0.9883	0.9962
DeepSeek-R1-32B	0.8409	0.8191	0.8376	0.9496	0.8622	0.9401	0.9862	0.9791	0.9865	0.9970	0.9931	0.9966
Llama-3.1-8B	0.8266	0.8101	0.8272	0.9058	0.8002	0.8885	0.9810	0.9746	0.9805	0.9961	0.9820	0.9957
Qwen3-8B	0.8273	0.8056	0.8217	0.9276	0.8045	0.9174	0.9820	0.9731	0.9817	0.9976	0.9899	0.9971
Phi-4	0.8311	0.8169	0.8365	0.9294	0.8673	0.9222	0.9838	0.9783	0.9837	0.9970	0.9910	0.9965

which is what Section 6.6 predicts from the other direction. Choosing where to read matters most when labels are scarce, and never hurts when they are not.

6.8 Cross-corpus transfer

A detector is only useful if it keeps working on data that differs from its training set. Table 15 trains on one corpus and tests on another. PARALLAX beats the last-layer probe in 58 of 60 pairs and 572 of 600 seeds, with a median gain of 0.10–0.18 per model. The last-layer probe scores at most 0.55 on 27 of the 60 pairs; PARALLAX never does. The two losses are small, and both come from probes trained on TruthfulQA. This comparison is against fixed-layer probes, which is what the rest of the paper contrasts with. A token-level reader transfers better than either. HaMI is ahead on 10 of 12 pairs on Llama-3.3-70B and 8 of 12 on DeepSeek-R1-32B (Appendix E.1). The cause is not the layer: HaMI's shallowest preset already lies inside the band PARALLAX selects, and giving it our selected layers changes its score by nothing. What is left is the unit each method scores. Pooling a response into one vector carries the surface characteristics of the source corpus along with the evidence, and those do not survive a change of corpus, whereas local evidence of an unsupported claim looks more alike across corpora. That is a different axis from the one studied here, and the two compose, since a token-level reader also has to choose a layer. It is not free either, costing about 140 times the activation storage per layer and several hundred labelled examples before it beats a pooled probe at all.

7 Discussion and Conclusion

Deployment. A new deployment may have few labelled examples and no reason to trust a choice made elsewhere. Three results bear on this. First, selection is cheap in labels. On Llama-3.3-70B RAGTruth, 100 labelled examples put the selected layer at 0.660 AUROC against 0.568 for the last layer. At 50 labelled examples it already leads, 0.595 against 0.531 (Section 6.6). Second, the choice is mostly a property of the model rather than the corpus, so it can be made once per model and reused. The chosen layers fall in a narrow band per model, layer 33 of 80 on Llama-3.3-70B in most folds, and the same bands reappear on 154 folds drawn from fresh splits (Appendix C.4). Third, where a corpus has no labels at all, layers selected on another still beat the fixed-layer probe in 58 of 60 transfers (Section 6.8).

Limitations: PARALLAX is supervised and its selection costs $O(k_{\max}L)$ probe fits on states extracted once. We study English corpora and open-weight models up to 70B, the two largest in 4-bit precision. Two boundaries are worth stating precisely. A token-level reader is stronger on RAGTruth at full supervision and transfers better between corpora (Appendix E.1), so our advantage there is confined to the low-label regime. And the sampling-based baselines, which need ten generations per question, were run on the two largest models only.

Conclusion. Evaluated on the text their labels describe, published hallucination detectors are weaker than their reported numbers suggest. Label-free internal scores sit near or below chance, and sampling-based uncertainty carries little signal once the text to be judged is supplied rather than generated. The layer is the variable that matters most. The informative ones sit in the middle of the network at a depth that moves between models (Gardinazzi et al., 2025), and inner cross-validation never picks the last layer in any of our 1,000 training folds. Choosing that layer on training data alone puts a plain logistic probe above every published single-pass detector we test, and wins all 20 settings against a fixed layer, with no sampled generations and nothing at inference beyond one forward pass and a linear classifier. Detectors that preselect a fixed layer leave much of this signal unexploited.

AI USE STATEMENT

Generative AI tools were used to help write evaluation code and to edit text drafted by the authors for grammatical errors and sentence flow. These tools were not used to generate research ideas, design experiments, write code from scratch without human supervision, or interpret results. All code was checked and tested by the authors against the reported outputs, and all text was reviewed by the authors. We take responsibility for the full content of this work.

ETHICS STATEMENT

This work evaluates hallucination detectors on public corpora under their licenses. It involves no human subjects and introduces no new data. Better detection can make LLM deployment safer; a detector can also give false confidence, which is why we report failure cases and transfer results.

REPRODUCIBILITY STATEMENT

Section 5 and Figure 3 define the protocol, Section 3 the method, and Appendices A.1–A.3 the data formats, baseline details, and compute. Anonymized code for the method, the layer-selection experiments, and the baseline re-evaluation is available at `https://anonymous.4open.science/r/llm-fuzzing-7613`, with a README mapping each table and figure to the script that produces it. All per-seed results will be released on publication.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. arXiv preprint arXiv:2304.13734.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv preprint arXiv:2212.03827.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv preprint arXiv:2402.03744.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv preprint arXiv:2309.03883.

Sharanya Dasgupta, Sujoy Nath, Arkaprabha Basu, Pourya Shamsolmoali, and Swagatam Das. HalluShift: Measuring distribution shifts towards hallucination detection in LLMs. *arXiv preprint arXiv:2504.09482*, 2025.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv preprint arXiv:2409.17504.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations (ICLR)*, 2025. Oral presentation. arXiv preprint arXiv:2411.14257.

Yuri Gardinazzi, Karthik Viswanathan, Giada Panerai, Alessio Ansuini, Alberto Cazzaniga, and Matteo Biagetti. Persistent topological features in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv preprint arXiv:2410.11042.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. Preprint at arXiv:2501.12948.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. doi: 10.1145/3703155. Preprint at arXiv:2311.05232.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in LLMs. *arXiv preprint arXiv:2508.08285*, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, Pascale Fung, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 10, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023a. Preprint at arXiv:2305.11747.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b. arXiv preprint arXiv:2306.03341.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3214–3252, 2022.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J.F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv preprint arXiv:2303.08896.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024. Preprint at arXiv:2310.06824.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 10862–10878, 2024.

Mengjia Niu, Hamed Haddadi, and Guansong Pang. Robust hallucination detection in LLMs via adaptive token selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Preprint at arXiv:2504.07863.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-time detection of hallucinated entities in long-form generation. *arXiv preprint arXiv:2509.03531*, 2025.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv preprint arXiv:2410.02707.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. *arXiv preprint arXiv:2502.14302*, 2025.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H Miller. Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Martin Preiß. Hallucination detection with the internal layers of LLMs. *arXiv preprint arXiv:2509.14254*, 2025.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach. In *Proceedings of the 26th International Conference on Artificial Intelligence (ICAI)*, 2024. Preprint at arXiv:2405.19648.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. Lynx: An open source hallucination evaluation model. *arXiv preprint arXiv:2407.08488*, 2024.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning (ICML)*, 2025. arXiv preprint arXiv:2502.02013.

Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. Unsupervised hallucination detection by inspecting reasoning processes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. arXiv preprint arXiv:2509.10004.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. arXiv preprint arXiv:2403.06448.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Han Li, et al. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free LLM truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM Web Conference (WWW)*, 2025a. arXiv preprint arXiv:2406.00034.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. Latent space chain-of-embedding enables output-free LLM self-evaluation. In *International Conference on Learning Representations (ICLR)*, 2025b. arXiv preprint arXiv:2410.13640.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. Prompt-guided internal states for hallucination detection of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025a. arXiv preprint arXiv:2411.04847.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. arXiv preprint arXiv:2305.13534.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023. GitHub: HillZhang1999/llm-hallucination-survey.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025b. arXiv preprint arXiv:2507.16488.

Zijian Zhang, Rizhen Hu, Athanasios Glentis, Dawei Li, Chung-Yiu Yau, Hongzhou Lin, and Mingyi Hong. Is one layer enough? training a single transformer layer can match full-parameter RL training. *arXiv preprint arXiv:2607.01232*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A EXPERIMENTAL DETAILS

Corpus formats, baseline definitions and adaptations, and compute.

A.1 CORPUS FORMATS

RAGTruth. The prompt is the task instruction with its source documents (news summarization, question answering, or data-to-text); the scored text is the stored response, truncated to 220 tokens (prompt: 3,500 tokens). **TruthfulQA.** `Question: {q}\n\nAnswer: {a}`, where a is one of the dataset's reference correct or incorrect answers. **HaluBench.** A passage, then `Question: {q}`, then `Answer: {a}`, labeled PASS/FAIL by the dataset. **MedHallu.** A context from PubMedQA, then the question, then `Answer: {a}`, where a is the ground-truth or the generated hallucinated answer. For the three candidate-answer corpora the scored text is a (up to 200 tokens; context up to 1,500 tokens).

A.2 BASELINES

Table 6 lists what each single-pass baseline reads and how it scores an example. All scores are computed on the labeled text described in Section 5. Supervised baselines use the outer splits of Section 5; any layer they choose is chosen inside the training fold. The fixed depths $60/70/80/85\%$ are the tap layers of the static multi-layer probe in Appendix C.2; the simplified baselines read the deepest of them. Label-free scores are evaluated on all 500 examples with the sign their method defines, so an AUROC below 0.5 means the score points the wrong way.

Table 6: Single-pass baselines. h_ℓ: layer-ℓ states averaged over the tokens of the scored text, as for PARALLAX; LR: standardized, class-balanced logistic regression (C=1).

Method	Score	Layer(s)	Labels
SAPLMA, last layer	LR on h_L	last	supervised
SAPLMA, 5 depths	LR or 256-128-64 MLP on h_ℓ	best of L, $L-L/8$, $L-L/4$, $L-3L/8$, $L/2$ by inner CV	supervised
SVD score (HaloScope-style)	reconstruction error of h_ℓ outside the top-64 SVD subspace of the corpus	85% depth	label-free
Two-layer cosine (HalluShift-style)	cosine distance between h_{ℓ_1} and h_{ℓ_2}	60% and 85% depth	label-free
Probe + prompt TF-IDF (PRISM-style)	LR on h_ℓ concatenated with a 128-d TF-IDF of the prompt	85% depth	supervised
Top-variance tokens (HaMI-style)	LR on the mean of the 8 highest-variance answer-token states	85% depth	supervised
Layer contrast (DoLa)	Jensen–Shannon divergence between the next-token distributions read out at two depths (early exit through the output head)	70% and 85% depth	label-free
LogProb	negative mean log-probability of the answer tokens	output	label-free
P(True)	$1 - P(\text{yes})/(P(\text{yes}) + P(\text{no}))$ for "Is the proposed answer factually correct and complete?"	output	label-free
Self-verification	$P(\text{No})/(P(\text{Yes}) + P(\text{No}))$ for "Is the proposed answer factually correct and well-grounded?"	output	label-free
TxTemb	LR on the MiniLM embedding of the scored text	none (no model access)	supervised

A.3 COMPUTE AND SOFTWARE

All runs use one server with eight NVIDIA L40S GPUs (46 GB) and 64 CPU cores. Hidden-state extraction uses one GPU per model with HuggingFace Transformers (4.57.6) and forward hooks on every decoder layer. Sampling uses vLLM with tensor parallelism. Baselines that need logits or prompting (DoLa, log-probability, P(True), self-verification) use one GPU forward pass per example. Probe training, layer selection, and HaMI training run on CPU with scikit-learn and PyTorch. Selection for one model and corpus takes 2 to 55 minutes on 16–40 cores, on states that

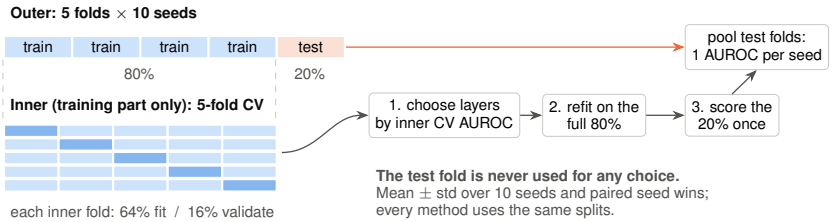

Figure 3: **Evaluation protocol.** Layer choice happens only inside the outer training set, by inner cross-validation. The chosen layers are refit on the whole training set and scored once on the held-out fold. Repeated for 5 folds and 10 seeds.

are extracted once. Inference is unchanged from a fixed-layer probe, one forward pass and a linear classifier.

B Uncertainty and metrics

AUROC over seeds measures the split. This adds the sample of examples, by bootstrap, and AUPRC for the imbalanced corpus.

B.1 AUPRC and bootstrap intervals

The seed spread in the main tables measures how much the result moves with the split, on a fixed set of 500 examples. Table 7 adds the other source of uncertainty by resampling examples. For each setting we draw 2,000 bootstrap resamples of the 500 examples, recompute pooled AUROC for both methods on each resample, average over the ten seeds, and report the 2.5th and 97.5th percentiles of the difference. The gain over the last-layer probe excludes zero in all 20 settings, with lower bounds from $+0.005$ to $+0.093$. The difference between the adaptive rule and one selected layer includes zero in 15 of the 20, which is the same conclusion the paired seed wins reach in Section 6.4.

The table also reports AUPRC, which matters on RAGTruth, where 153 of 500 examples are hallucinated. There AUROC near 0.75 corresponds to AUPRC between 0.48 and 0.60 against a base rate of 0.31, so the ranking of methods is unchanged but the absolute numbers are far less flattering than AUROC alone suggests. Neither metric speaks to whether a detector's scores are calibrated, which is a separate question with its own literature (Guo et al., 2017).

Table 7: AUPRC and 95% bootstrap intervals over examples (2,000 resamples). Differences are in pooled AUROC.

Model	Corpus	PARALLAX AUPRC	last-layer AUPRC	Δ vs last layer (95% CI)	Δ vs k=1 (95% CI)
Llama-3.3-70B	RAGTruth	0.6006	0.4838	+0.089 [+0.059, +0.117]	+0.016 [+0.007, +0.026]
Llama-3.3-70B	TruthfulQA	0.9551	0.8528	+0.089 [+0.068, +0.112]	-0.001 [-0.003, -0.000]
Llama-3.3-70B	HaluBench	0.9578	0.9309	+0.029 [+0.016, +0.044]	+0.004 [-0.001, +0.009]
Llama-3.3-70B	MedHallu	0.9918	0.9702	+0.019 [+0.011, +0.028]	+0.001 [-0.000, +0.002]
DeepSeek-R1-32B	RAGTruth	0.5845	0.4528	+0.096 [+0.067, +0.125]	-0.003 [-0.011, +0.006]
DeepSeek-R1-32B	TruthfulQA	0.9422	0.8503	+0.088 [+0.065, +0.111]	-0.003 [-0.006, +0.000]
DeepSeek-R1-32B	HaluBench	0.9629	0.9132	+0.053 [+0.037, +0.072]	-0.002 [-0.005, +0.000]
DeepSeek-R1-32B	MedHallu	0.9953	0.9794	+0.015 [+0.008, +0.022]	+0.000 [-0.001, +0.002]
Llama-3.1-8B	RAGTruth	0.4812	0.4211	+0.064 [+0.035, +0.093]	+0.020 [+0.011, +0.030]
Llama-3.1-8B	TruthfulQA	0.9121	0.7760	+0.118 [+0.093, +0.145]	+0.001 [-0.003, +0.004]
Llama-3.1-8B	HaluBench	0.9536	0.9204	+0.032 [+0.017, +0.048]	-0.001 [-0.004, +0.002]
Llama-3.1-8B	MedHallu	0.9897	0.9535	+0.033 [+0.021, +0.047]	-0.001 [-0.001, -0.000]
Qwen3-8B	RAGTruth	0.5587	0.4401	+0.107 [+0.076, +0.139]	+0.019 [+0.011, +0.028]
Qwen3-8B	TruthfulQA	0.9098	0.7785	+0.111 [+0.085, +0.137]	-0.005 [-0.009, -0.001]
Qwen3-8B	HaluBench	0.9522	0.9175	+0.041 [+0.026, +0.058]	-0.002 [-0.004, +0.001]
Qwen3-8B	MedHallu	0.9929	0.9740	+0.020 [+0.013, +0.029]	-0.000 [-0.001, +0.000]
Phi-4	RAGTruth	0.4797	0.4490	+0.031 [+0.007, +0.056]	+0.008 [-0.000, +0.015]
Phi-4	TruthfulQA	0.9171	0.8573	+0.054 [+0.036, +0.072]	+0.000 [-0.003, +0.003]
Phi-4	HaluBench	0.9542	0.9303	+0.027 [+0.013, +0.042]	+0.002 [-0.001, +0.005]
Phi-4	MedHallu	0.9930	0.9797	+0.012 [+0.005, +0.021]	-0.001 [-0.002, -0.000]

C Results per setting

Every cell behind the averages in the main text.

C.1 PER MODEL AND CORPUS

One table per corpus, behind the five-model averages in Table 2. The spread between the best and worst model for one detector reaches 0.18 AUROC on RAGTruth, wider than most gaps between detectors on the same model.

Table 8: RAGTruth: AUROC per model (mean over 10 seeds for supervised methods).

Method	Llama-3.3-70B	DeepSeek-R1-32B	Llama-3.1-8B	Qwen3-8B	Phi-4
TXTEMB	0.523	0.523	0.523	0.523	0.523
SAPLMA, last layer	0.658	0.642	0.610	0.626	0.639
SAPLMA, 5 depths (LR)	0.675	0.720	0.660	0.707	0.668
SAPLMA, 5 depths (MLP)	0.636	0.663	0.632	0.669	0.644
HaloScope	0.514	0.518	0.537	0.500	0.514
HalluShift	0.668	0.643	0.622	0.585	0.640
PRISM	0.691	0.660	0.575	0.627	0.561
Layer contrast (DoLa)†	0.550	0.473	0.585	0.506	0.498
LogProb†	0.635	0.631	0.588	0.619	0.647
P(True)†	0.566	0.730	0.474	0.718	0.478
Self-verification†	0.616	0.729	0.582	0.698	0.548
PARALLAX	0.746	0.738	0.673	0.733	0.670

Table 9: TruthfulQA: AUROC per model (mean over 10 seeds for supervised methods).

Method	Llama-3.3-70B	DeepSeek-R1-32B	Llama-3.1-8B	Qwen3-8B	Phi-4
TXTEMB	0.647	0.647	0.647	0.647	0.647
SAPLMA, last layer	0.864	0.855	0.795	0.799	0.862
SAPLMA, 5 depths (LR)	0.934	0.939	0.911	0.896	0.912
SAPLMA, 5 depths (MLP)	0.917	0.930	0.900	0.905	0.904
HaloScope	0.582	0.544	0.573	0.528	0.571
HalluShift	0.735	0.652	0.654	0.675	0.696
PRISM	0.927	0.919	0.863	0.836	0.876
Layer contrast (DoLa)†	0.550	0.399	0.514	0.533	0.443
LogProb†	0.584	0.536	0.557	0.554	0.568
P(True)†	0.806	0.796	0.580	0.760	0.538
Self-verification†	0.866	0.775	0.687	0.784	0.716
PARALLAX	0.953	0.943	0.913	0.910	0.915

Table 10: HaluBench: AUROC per model (mean over 10 seeds for supervised methods).

Method	Llama-3.3-70B	DeepSeek-R1-32B	Llama-3.1-8B	Qwen3-8B	Phi-4
TXTEMB	0.789	0.789	0.789	0.789	0.789
SAPLMA, last layer	0.925	0.907	0.915	0.907	0.921
SAPLMA, 5 depths (LR)	0.945	0.959	0.945	0.946	0.944
SAPLMA, 5 depths (MLP)	0.949	0.957	0.950	0.949	0.947
HaloScope	0.774	0.666	0.688	0.623	0.721
HalluShift	0.905	0.872	0.868	0.837	0.857
PRISM	0.938	0.947	0.926	0.940	0.937
Layer contrast (DoLa)†	0.721	0.297	0.655	0.588	0.507
LogProb†	0.481	0.491	0.462	0.475	0.439
P(True)†	0.886	0.885	0.656	0.892	0.509
Self-verification†	0.894	0.909	0.857	0.875	0.860
PARALLAX	0.954	0.960	0.948	0.948	0.948

C.2 VARIANTS OF THE METHOD

Best single is the layer chosen by inner CV, *last layer* the SAPLMA-style probe, *fixed four* the taps at 60, 70, 80 and 85% depth, and *depth average* the mean over all layers. Choosing one layer beats every fixed alternative in all 20 settings, on all 10 seeds.

Table 11: MedHallu: AUROC per model (mean over 10 seeds for supervised methods).

Method	Llama-3.3-70B	DeepSeek-R1-32B	Llama-3.1-8B	Qwen3-8B	Phi-4
TXTEMB	0.676	0.676	0.676	0.676	0.676
SAPLMA, last layer	0.975	0.981	0.957	0.973	0.980
SAPLMA, 5 depths (LR)	0.993	0.995	0.989	0.994	0.994
SAPLMA, 5 depths (MLP)	0.989	0.993	0.983	0.991	0.993
HaloScope	0.770	0.692	0.815	0.711	0.807
HalluShift	0.946	0.935	0.888	0.842	0.944
PRISM	0.986	0.986	0.977	0.978	0.996
Layer contrast (DoLa)†	0.591	0.422	0.634	0.491	0.394
LogProb†	0.407	0.259	0.370	0.324	0.354
P(True)†	0.875	0.964	0.650	0.944	0.478
Self-verification†	0.936	0.957	0.893	0.927	0.959
PARALLAX	0.994	0.995	0.990	0.993	0.992

Table 12: PARALLAX (adaptive), the best single layer chosen on training folds, the last-layer probe, a probe on four fixed layers (60/70/80/85% depth), and a probe on the average over all layers. Last column: PARALLAX minus last layer (seeds won).

Model	Corpus	PARALLAX	Best single	Last layer	Fixed 4 layers	All-layer mean	Gain (wins)
Llama-3.3-70B	RAGTruth	0.746 ± 0.015	0.730	0.658	0.677	0.673	+0.088 (10/10)
Llama-3.3-70B	TruthfulQA	0.953 ± 0.005	0.955	0.864	0.846	0.912	+0.089 (10/10)
Llama-3.3-70B	HaluBench	0.954 ± 0.003	0.951	0.925	0.945	0.948	+0.029 (10/10)
Llama-3.3-70B	MedHallu	0.994 ± 0.001	0.993	0.975	0.975	0.990	+0.019 (10/10)
DeepSeek-R1-32B	RAGTruth	0.738 ± 0.014	0.741	0.642	0.690	0.649	+0.096 (10/10)
DeepSeek-R1-32B	TruthfulQA	0.943 ± 0.004	0.946	0.855	0.895	0.890	+0.088 (10/10)
DeepSeek-R1-32B	HaluBench	0.960 ± 0.004	0.962	0.907	0.951	0.942	+0.053 (10/10)
DeepSeek-R1-32B	MedHallu	0.995 ± 0.001	0.995	0.981	0.989	0.990	+0.015 (10/10)
Llama-3.1-8B	RAGTruth	0.673 ± 0.018	0.653	0.610	0.639	0.624	+0.063 (10/10)
Llama-3.1-8B	TruthfulQA	0.913 ± 0.005	0.912	0.795	0.791	0.842	+0.118 (10/10)
Llama-3.1-8B	HaluBench	0.948 ± 0.006	0.949	0.915	0.935	0.934	+0.032 (10/10)
Llama-3.1-8B	MedHallu	0.990 ± 0.001	0.990	0.957	0.967	0.980	+0.033 (10/10)
Qwen3-8B	RAGTruth	0.733 ± 0.014	0.714	0.626	0.648	0.649	+0.107 (10/10)
Qwen3-8B	TruthfulQA	0.910 ± 0.007	0.914	0.799	0.806	0.836	+0.111 (10/10)
Qwen3-8B	HaluBench	0.948 ± 0.005	0.950	0.907	0.934	0.936	+0.041 (10/10)
Qwen3-8B	MedHallu	0.993 ± 0.001	0.994	0.973	0.977	0.986	+0.020 (10/10)
Phi-4	RAGTruth	0.670 ± 0.015	0.663	0.639	0.646	0.634	+0.031 (10/10)
Phi-4	TruthfulQA	0.915 ± 0.006	0.915	0.862	0.832	0.892	+0.053 (10/10)
Phi-4	HaluBench	0.948 ± 0.003	0.946	0.921	0.928	0.931	+0.027 (10/10)
Phi-4	MedHallu	0.992 ± 0.001	0.993	0.980	0.969	0.990	+0.012 (10/10)

C.3 EXACTLY k LAYERS

Table 13 gives every k per model and corpus, with the range over the ten seeds and the layers the selection reaches for most often. The selected layers are concentrated in a narrow band per model, and that band sits in the middle of the network in every case. With the adaptive rule PARALLAX uses 2.6 layers on average.

AUROC is not monotone in k, because the second layer is chosen on inner-CV score and need not help on held-out data. This is most visible on DeepSeek-R1-32B RAGTruth, 0.741 at $k = 1$, 0.731 at $k = 2$ and 0.741 again at $k = 4$.

Choosing k on training data changes nothing, and neither does tuning the probe. Picking k per fold by inner-CV AUROC gives 0.713, 0.927, 0.952 and 0.993, within 0.001 of the adaptive rule everywhere, and selecting the regularization strength per fold from $\{0.01, 0.1, 1, 10\}$ is worse than the fixed $C{=}1$ in most settings, for example 0.936 against 0.944 on DeepSeek-R1-32B TruthfulQA with 0 of 10 seeds won. What the probe reads matters more than its regularization or its depth.

C.4 FULL CORPUS SIZE

Every cell with its seed standard deviation, paired split wins, and the layers returned. The spread is 0.0006 to 0.0078 on the large corpora and up to 0.0096 on TruthfulQA, which has only 1,634 examples. The layers match the bands of the 500-example runs in Table 13.

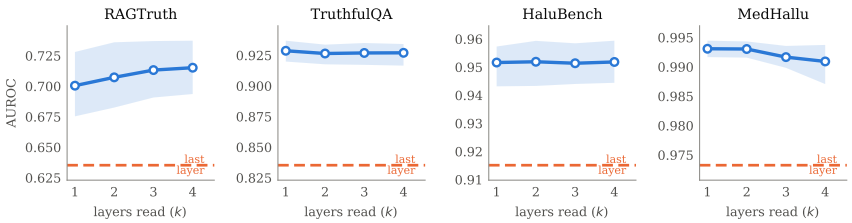

Figure 4: **How many layers to read.** AUROC against the number of selected layers, averaged over the five models, with the band spanning the best and worst seed. The dashed line is the last-layer probe. Extra layers pay off only on RAGTruth, the one corpus whose labelled text is a long free-form response; every setting of k is far above the fixed layer.

Table 13: Exactly k layers, per model and corpus. Mean AUROC with [min, max] over the ten seeds, the four most frequently selected layers at $k = 4$, and the model's depth.

Model	Corpus	$k=1$	$k=2$	$k=3$	$k=4$	most selected	L
Llama-3.3-70B	RAGTruth	0.730 [0.700,0.753]	0.743 [0.703,0.764]	0.750 [0.712,0.771]	0.749 [0.721,0.765]	33, 32, 31, 34	80
Llama-3.3-70B	TruthfulQA	0.955 [0.945,0.963]	0.954 [0.945,0.961]	0.946 [0.934,0.957]	0.949 [0.935,0.958]	32, 34, 31, 33	80
Llama-3.3-70B	HaluBench	0.951 [0.941,0.954]	0.956 [0.949,0.963]	0.955 [0.949,0.963]	0.953 [0.947,0.962]	29, 32, 33, 34	80
Llama-3.3-70B	MedHallu	0.993 [0.991,0.995]	0.994 [0.993,0.996]	0.994 [0.991,0.996]	0.992 [0.987,0.995]	33, 34, 31, 32	80
DeepSeek-R1-32B	RAGTruth	0.741 [0.717,0.778]	0.731 [0.716,0.767]	0.739 [0.726,0.762]	0.741 [0.726,0.766]	42, 43, 45, 41	64
DeepSeek-R1-32B	TruthfulQA	0.946 [0.940,0.952]	0.944 [0.937,0.951]	0.943 [0.936,0.951]	0.943 [0.937,0.950]	43, 45, 44, 42	64
DeepSeek-R1-32B	HaluBench	0.962 [0.957,0.967]	0.960 [0.954,0.966]	0.959 [0.953,0.964]	0.960 [0.952,0.966]	45, 41, 43, 44	64
DeepSeek-R1-32B	MedHallu	0.995 [0.994,0.996]	0.995 [0.994,0.996]	0.995 [0.993,0.996]	0.994 [0.992,0.996]	43, 44, 41, 42	64
Llama-3.1-8B	RAGTruth	0.653 [0.627,0.681]	0.663 [0.642,0.701]	0.673 [0.654,0.706]	0.676 [0.657,0.705]	13, 15, 14, 11	32
Llama-3.1-8B	TruthfulQA	0.912 [0.902,0.924]	0.912 [0.903,0.918]	0.915 [0.906,0.921]	0.915 [0.903,0.922]	14, 11, 12, 0	32
Llama-3.1-8B	HaluBench	0.949 [0.934,0.956]	0.947 [0.930,0.957]	0.948 [0.934,0.955]	0.950 [0.938,0.957]	13, 14, 2, 12	32
Llama-3.1-8B	MedHallu	0.990 [0.989,0.992]	0.990 [0.987,0.991]	0.987 [0.984,0.989]	0.988 [0.980,0.993]	14, 12, 13, 1	32
Qwen3-8B	RAGTruth	0.714 [0.693,0.739]	0.730 [0.711,0.754]	0.731 [0.712,0.750]	0.732 [0.709,0.752]	21, 14, 15, 20	36
Qwen3-8B	TruthfulQA	0.914 [0.903,0.922]	0.910 [0.898,0.918]	0.911 [0.901,0.920]	0.909 [0.896,0.919]	21, 15, 20, 18	36
Qwen3-8B	HaluBench	0.950 [0.943,0.958]	0.950 [0.945,0.960]	0.947 [0.941,0.956]	0.949 [0.941,0.955]	20, 7, 19, 5	36
Qwen3-8B	MedHallu	0.994 [0.992,0.995]	0.993 [0.992,0.995]	0.992 [0.991,0.993]	0.991 [0.990,0.993]	21, 0, 19, 22	36
Phi-4	RAGTruth	0.663 [0.638,0.689]	0.667 [0.640,0.693]	0.672 [0.649,0.696]	0.675 [0.654,0.697]	18, 16, 21, 11	40
Phi-4	TruthfulQA	0.915 [0.909,0.924]	0.912 [0.903,0.919]	0.918 [0.908,0.925]	0.917 [0.910,0.921]	23, 17, 0, 15	40
Phi-4	HaluBench	0.946 [0.940,0.951]	0.946 [0.937,0.952]	0.948 [0.943,0.954]	0.948 [0.943,0.956]	17, 16, 6, 21	40
Phi-4	MedHallu	0.993 [0.992,0.994]	0.993 [0.991,0.994]	0.991 [0.990,0.993]	0.989 [0.985,0.991]	18, 15, 16, 0	40

C.5 Cross-corpus transfer

Figure 5 shows the gain of PARALLAX over the last-layer probe for every train/test pair; the tables below give the underlying AUROCs. Layers are chosen and the probe is trained on a stratified 80% of the training corpus, and the whole test corpus is scored (10 seeds).

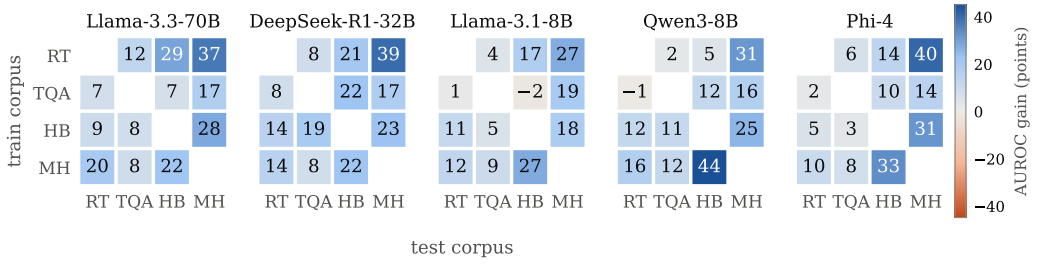

Figure 5: Cross-corpus transfer: PARALLAX minus last-layer SAPLMA, in AUROC points (mean over 10 seeds). Rows: training corpus; columns: test corpus. RT: RAGTruth, TQA: TruthfulQA, HB: HaluBench, MH: MedHallu. Blue: PARALLAX better.

D Ablations on the candidate set

How many candidate depths the search needs, and how far a simplified version of each published method gets.

D.1 Candidate-set size

SAPLMA's five depths differ from a full search both in number and in placement. To separate the two, for $m \in \{5, 10, 20\}$ we search m evenly spaced depths from $L/4$ to the output, pick one by

Table 14: Full corpus size, per setting. PARALLAX (k=1), the last-layer probe and SAPLMA's five depths, mean ± standard deviation over five seeds, with PARALLAX's paired split wins and the selected layers.

Model	Corpus	n	PARALLAX	Last layer	SAPLMA-5	vs last	vs S-5	layers
DeepSeek-R1-32B	HaluBench	14340	0.9862 ± 0.0006	0.9791 ± 0.0019	0.9865 ± 0.0015	+0.0071 (5/5)	-0.0002 (2/5)	43, 44, 45, 47
DeepSeek-R1-32B	MedHallu	2000	0.9970 ± 0.0013	0.9931 ± 0.0019	0.9966 ± 0.0015	+0.0039 (5/5)	+0.0005 (3/5)	43, 46, 48
DeepSeek-R1-32B	RAGTruth	15090	0.8409 ± 0.0038	0.8191 ± 0.0042	0.8376 ± 0.0035	+0.0219 (5/5)	+0.0034 (5/5)	43, 44, 45
DeepSeek-R1-32B	TruthfulQA	1634	0.9496 ± 0.0033	0.8622 ± 0.0117	0.9401 ± 0.0058	+0.0874 (5/5)	+0.0096 (5/5)	43, 44
Llama-3.1-8B	HaluBench	14340	0.9810 ± 0.0013	0.9746 ± 0.0015	0.9805 ± 0.0010	+0.0065 (5/5)	+0.0006 (3/5)	12, 13, 14, 16
Llama-3.1-8B	MedHallu	2000	0.9961 ± 0.0015	0.9820 ± 0.0042	0.9957 ± 0.0015	+0.0141 (5/5)	+0.0004 (5/5)	12, 13, 14
Llama-3.1-8B	RAGTruth	15090	0.8266 ± 0.0061	0.8101 ± 0.0075	0.8272 ± 0.0069	+0.0165 (5/5)	-0.0006 (2/5)	14, 15
Llama-3.1-8B	TruthfulQA	1634	0.9058 ± 0.0096	0.8002 ± 0.0148	0.8885 ± 0.0054	+0.1056 (5/5)	+0.0172 (5/5)	12, 13
Llama-3.3-70B	HaluBench	14340	0.9880 ± 0.0016	0.9820 ± 0.0013	0.9869 ± 0.0010	+0.0059 (5/5)	+0.0010 (4/5)	34, 70
Llama-3.3-70B	MedHallu	2000	0.9969 ± 0.0008	0.9883 ± 0.0050	0.9962 ± 0.0010	+0.0086 (5/5)	+0.0006 (3/5)	28, 32, 33
Llama-3.3-70B	RAGTruth	15090	0.8559 ± 0.0067	0.8361 ± 0.0027	0.8517 ± 0.0035	+0.0198 (5/5)	+0.0042 (3/5)	32, 33
Llama-3.3-70B	TruthfulQA	1634	0.9524 ± 0.0039	0.8881 ± 0.0096	0.9327 ± 0.0023	+0.0644 (5/5)	+0.0197 (5/5)	31, 32
Phi-4	HaluBench	14340	0.9838 ± 0.0013	0.9783 ± 0.0010	0.9837 ± 0.0014	+0.0055 (5/5)	+0.0001 (2/5)	19, 20, 24
Phi-4	MedHallu	2000	0.9970 ± 0.0009	0.9910 ± 0.0015	0.9965 ± 0.0018	+0.0060 (5/5)	+0.0005 (4/5)	16, 17, 21
Phi-4	RAGTruth	15090	0.8311 ± 0.0072	0.8169 ± 0.0100	0.8365 ± 0.0043	+0.0142 (5/5)	-0.0054 (1/5)	18, 19, 20
Phi-4	TruthfulQA	1634	0.9294 ± 0.0060	0.8673 ± 0.0040	0.9222 ± 0.0069	+0.0621 (5/5)	+0.0071 (5/5)	20, 21
Qwen3-8B	HaluBench	14340	0.9820 ± 0.0010	0.9731 ± 0.0016	0.9817 ± 0.0017	+0.0089 (5/5)	+0.0003 (2/5)	20, 21, 22, 23
Qwen3-8B	MedHallu	2000	0.9976 ± 0.0007	0.9899 ± 0.0033	0.9971 ± 0.0011	+0.0077 (5/5)	+0.0005 (4/5)	21
Qwen3-8B	RAGTruth	15090	0.8273 ± 0.0078	0.8056 ± 0.0057	0.8217 ± 0.0058	+0.0218 (5/5)	+0.0057 (4/5)	21, 23
Qwen3-8B	TruthfulQA	1634	0.9276 ± 0.0059	0.8045 ± 0.0128	0.9174 ± 0.0077	+0.1231 (5/5)	+0.0102 (5/5)	20, 21

Table 15: **Cross-corpus transfer.** Layers selected and probe trained on 80% of one corpus, tested on all of another (12 ordered pairs per model, 10 seeds). Pairs and seeds won against the last-layer probe, median and largest gain, and the number of pairs where each method scores at most 0.55.

Model	Pairs won	Seeds won	Median gain	Largest gain	Last layer ≤ 0.55	PARALLAX ≤ 0.55
Llama-3.3-70B	12/12	119/120	+0.148	+0.366	4	0
DeepSeek-R1-32B	12/12	120/120	+0.181	+0.391	5	0
Llama-3.1-8B	11/12	109/120	+0.116	+0.274	6	0
Qwen3-8B	11/12	107/120	+0.118	+0.444	6	0
Phi-4	12/12	117/120	+0.099	+0.404	6	0

inner CV on the training fold as PARALLAX does, and score the held-out fold. The last column searches all L layers.

Averaged over the 20 settings, the full search is ahead of m=5 by $+0.005$ and leads in 13 of 20 settings, ahead of m=10 by $+0.001$ and leads in 8, and behind m=20 by -0.001, leading in 6. Ten well-placed candidates therefore recover almost all of what searching every layer provides. A five-point grid over the same range comes within $+0.005$, while SAPLMA's five preset depths lose by up to $+0.071$ (Table 3), so the gain comes from covering the middle of the network rather than from the number of candidates. A practitioner can have it at a tenth of the search cost if the grid is placed there.

D.2 Simplified reimplementations

Before the released implementations were available we scored simplified versions of HaloScope, HalluShift, PRISM and HaMI inside our own harness, each keeping the core feature of the method. Table 22 puts the two side by side, averaged over the five models. The simplified scores are not substitutes, and the direction of the error is not consistent, so only the official runs appear in the main table.

E Detectors that read tokens or samples

The families that do not reduce the scored text to a single pooled vector.

E.1 A token-level reader

HaMI (Niu et al., 2025) scores each token with a multiple-instance network over four preset layers, rather than pooling the scored text into one vector. On RAGTruth at full supervision it is the strongest detector we run, 0.758 on Llama-3.3-70B and 0.764 on DeepSeek-R1-32B against 0.746 and 0.739 for PARALLAX, and on the other three corpora PARALLAX is ahead by 0.011 to 0.036. Three further results place that advantage.

It does not come from where it reads. Replacing HaMI's preset layers with the layers PARALLAX selects on the same training fold changes nothing, whether we give it the single best layer or the whole selected set. Its shallowest preset, at $0.4L$, already falls inside the band our selection prefers, layers 29–35 of 80 and 42–46 of 64, so there is nothing for selection to recover.

Table 16: Cross-corpus transfer, Llama-3.3-70B (mean over 10 seeds).

Train → test	PARALLAX	Best single	Last layer	Fixed 4 layers	Gain
TruthfulQA → RAGTruth	0.644	0.644	0.571	0.571	+0.072
HaluBench → RAGTruth	0.668	0.668	0.574	0.634	+0.094
MedHallu → RAGTruth	0.709	0.697	0.513	0.587	+0.196
RAGTruth → TruthfulQA	0.777	0.777	0.653	0.676	+0.125
HaluBench → TruthfulQA	0.695	0.703	0.616	0.647	+0.079
MedHallu → TruthfulQA	0.755	0.764	0.670	0.648	+0.085
RAGTruth → HaluBench	0.784	0.787	0.496	0.595	+0.288
TruthfulQA → HaluBench	0.713	0.710	0.645	0.460	+0.068
MedHallu → HaluBench	0.742	0.738	0.526	0.523	+0.216
RAGTruth → MedHallu	0.906	0.906	0.540	0.633	+0.366
TruthfulQA → MedHallu	0.896	0.898	0.725	0.627	+0.171
HaluBench → MedHallu	0.830	0.832	0.555	0.493	+0.275

Table 17: Cross-corpus transfer, DeepSeek-R1-32B (mean over 10 seeds).

Train → test	PARALLAX	Best single	Last layer	Fixed 4 layers	Gain
TruthfulQA → RAGTruth	0.665	0.682	0.587	0.627	+0.078
HaluBench → RAGTruth	0.676	0.673	0.533	0.611	+0.143
MedHallu → RAGTruth	0.720	0.724	0.576	0.655	+0.145
RAGTruth → TruthfulQA	0.741	0.745	0.659	0.697	+0.082
HaluBench → TruthfulQA	0.719	0.723	0.532	0.636	+0.187
MedHallu → TruthfulQA	0.729	0.758	0.651	0.700	+0.078
RAGTruth → HaluBench	0.686	0.679	0.481	0.570	+0.206
TruthfulQA → HaluBench	0.790	0.801	0.567	0.632	+0.224
MedHallu → HaluBench	0.753	0.791	0.538	0.748	+0.215
RAGTruth → MedHallu	0.900	0.888	0.509	0.830	+0.391
TruthfulQA → MedHallu	0.845	0.871	0.671	0.801	+0.174
HaluBench → MedHallu	0.808	0.799	0.581	0.727	+0.227

It does not come from sampling. HaMI weights token bags by the agreement of ten sampled answers. Setting every weight to one changes its score by $+0.001$ on Llama-3.3-70B and -0.007 on DeepSeek-R1-32B, so the method is effectively single-pass and we report it as such.

It costs labels. Table 23 varies the number of labelled examples in each training fold. Below 200 labels PARALLAX wins every seed on both models, by 0.048 to 0.071 AUROC, and at 50 labels HaMI is close to chance. The crossover sits between 200 and 400 labels on both models.

Set against this, a token-level reader transfers better. Table 24 trains on one corpus and scores another. HaMI is ahead on 10 of 12 pairs for Llama-3.3-70B and 8 of 12 for DeepSeek-R1-32B, with the largest gaps on transfers into HaluBench. The two designs trade off, and the trade is legible. Reading tokens buys robustness to a change of corpus and costs roughly 140 times the activation storage per layer, 100 epochs of network training per fold, and several hundred labelled examples before it pays for itself.

E.2 Sampling-based detectors

These detectors draw ten samples per query and score the labelled text against them.

E.3 Comparing across implementations

One caution applies to anyone comparing these detectors. Reimplementing a multiple-instance baseline while changing only where its random stream is created moved its AUROC by 0.011 on one model. Candidates that differ in batch order as well as in layer let the validation split reward luck that does not transfer to the test fold, so the effect grows with the number of candidates a method compares. Our first attempt at combining layer selection with that reader appeared to win by 0.005 for exactly this reason and did not survive a like-for-like rerun. Numbers from two implementations of the same method should not be compared, and every comparison we report is within-implementation.

Table 18: Cross-corpus transfer, Llama-3.1-8B (mean over 10 seeds).

Train → test	PARALLAX	Best single	Last layer	Fixed 4 layers	Gain
TruthfulQA → RAGTruth	0.597	0.592	0.585	0.586	+0.013
HaluBench → RAGTruth	0.655	0.662	0.542	0.625	+0.112
MedHallu → RAGTruth	0.661	0.659	0.541	0.584	+0.120
RAGTruth → TruthfulQA	0.617	0.643	0.578	0.600	+0.039
HaluBench → TruthfulQA	0.573	0.590	0.523	0.544	+0.050
MedHallu → TruthfulQA	0.690	0.690	0.599	0.626	+0.091
RAGTruth → HaluBench	0.707	0.707	0.538	0.527	+0.169
TruthfulQA → HaluBench	0.564	0.568	0.584	0.558	-0.020
MedHallu → HaluBench	0.643	0.643	0.369	0.478	+0.274
RAGTruth → MedHallu	0.823	0.835	0.556	0.518	+0.268
TruthfulQA → MedHallu	0.799	0.803	0.608	0.665	+0.191
HaluBench → MedHallu	0.666	0.683	0.486	0.462	+0.180

Table 19: Cross-corpus transfer, Qwen3-8B (mean over 10 seeds).

Train → test	PARALLAX	Best single	Last layer	Fixed 4 layers	Gain
TruthfulQA → RAGTruth	0.596	0.593	0.605	0.588	-0.009
HaluBench → RAGTruth	0.656	0.644	0.539	0.567	+0.117
MedHallu → RAGTruth	0.705	0.701	0.541	0.587	+0.164
RAGTruth → TruthfulQA	0.650	0.624	0.627	0.604	+0.024
HaluBench → TruthfulQA	0.613	0.607	0.503	0.530	+0.110
MedHallu → TruthfulQA	0.712	0.702	0.594	0.614	+0.118
RAGTruth → HaluBench	0.624	0.604	0.579	0.516	+0.045
TruthfulQA → HaluBench	0.682	0.684	0.565	0.423	+0.117
MedHallu → HaluBench	0.773	0.788	0.329	0.503	+0.444
RAGTruth → MedHallu	0.862	0.846	0.550	0.630	+0.312
TruthfulQA → MedHallu	0.804	0.790	0.640	0.645	+0.164
HaluBench → MedHallu	0.738	0.742	0.487	0.542	+0.251

Table 20: Cross-corpus transfer, Phi-4 (mean over 10 seeds).

Train → test	PARALLAX	Best single	Last layer	Fixed 4 layers	Gain
TruthfulQA → RAGTruth	0.604	0.596	0.586	0.584	+0.018
HaluBench → RAGTruth	0.648	0.638	0.596	0.630	+0.052
MedHallu → RAGTruth	0.635	0.649	0.536	0.512	+0.100
RAGTruth → TruthfulQA	0.707	0.699	0.646	0.660	+0.061
HaluBench → TruthfulQA	0.620	0.592	0.593	0.580	+0.027
MedHallu → TruthfulQA	0.654	0.651	0.578	0.524	+0.075
RAGTruth → HaluBench	0.627	0.618	0.491	0.458	+0.136
TruthfulQA → HaluBench	0.627	0.618	0.528	0.402	+0.098
MedHallu → HaluBench	0.669	0.667	0.341	0.335	+0.327
RAGTruth → MedHallu	0.835	0.849	0.431	0.428	+0.404
TruthfulQA → MedHallu	0.792	0.773	0.648	0.601	+0.144
HaluBench → MedHallu	0.691	0.661	0.377	0.377	+0.314

Table 21: Candidate-set size. AUROC when the search is restricted to m evenly spaced depths from $L/4$ to the output, against searching all L layers, under the protocol of Section 5.

Model	Corpus	m=5	m=10	m=20	all L
Llama-3.3-70B	RAGTruth	0.723	0.734	0.726	0.730
Llama-3.3-70B	TruthfulQA	0.949	0.955	0.957	0.955
Llama-3.3-70B	HaluBench	0.945	0.952	0.953	0.951
Llama-3.3-70B	MedHallu	0.993	0.994	0.993	0.993
DeepSeek-R1-32B	RAGTruth	0.723	0.746	0.744	0.741
DeepSeek-R1-32B	TruthfulQA	0.914	0.938	0.947	0.946
DeepSeek-R1-32B	HaluBench	0.945	0.958	0.962	0.962
DeepSeek-R1-32B	MedHallu	0.992	0.995	0.995	0.995
Llama-3.1-8B	RAGTruth	0.664	0.653	0.656	0.653
Llama-3.1-8B	TruthfulQA	0.918	0.909	0.913	0.912
Llama-3.1-8B	HaluBench	0.951	0.951	0.949	0.949
Llama-3.1-8B	MedHallu	0.991	0.989	0.991	0.990
Qwen3-8B	RAGTruth	0.708	0.719	0.717	0.714
Qwen3-8B	TruthfulQA	0.898	0.915	0.914	0.914
Qwen3-8B	HaluBench	0.946	0.945	0.951	0.950
Qwen3-8B	MedHallu	0.994	0.994	0.994	0.994
Phi-4	RAGTruth	0.652	0.651	0.664	0.663
Phi-4	TruthfulQA	0.918	0.915	0.913	0.915
Phi-4	HaluBench	0.948	0.947	0.946	0.946
Phi-4	MedHallu	0.993	0.993	0.993	0.993

Table 22: Simplified versions from our harness against the released code, mean AUROC over the five models. HaMI has no official run outside RAGTruth.

	RAGTruth		TruthfulQA		HaluBench		MedHallu	
Method	ours	official	ours	official	ours	official	ours	official
SVD score (HaloScope-style)	0.560	0.517	0.613	0.559	0.435	0.694	0.485	0.759
Two-layer cosine (HalluShift-style)	0.507	0.631	0.502	0.682	0.334	0.868	0.382	0.911
Probe + prompt TF-IDF (PRISM-style)	0.641	0.623	0.850	0.884	0.930	0.937	0.983	0.984
Top-variance tokens (HaMI-style)	0.612	–	0.842	–	0.932	–	0.973	–

Table 23: Label efficiency on RAGTruth. AUROC with the training fold subsampled to 50, 100, 200 and 400 labelled examples. "last" is the same probe read from the last layer, a control for whether the advantage comes from the layer or from the classifier.

	50 labels			100 labels			200 labels			400 labels		
Model	ours	last	HaMI	ours	last	HaMI	ours	last	HaMI	ours	last	HaMI
Llama-3.3-70B	0.595	0.547	10/10	0.660	0.589	10/10	0.700	0.668	10/10	0.746	0.758	2/10
DeepSeek-R1-32B	0.597	0.537	10/10	0.659	0.593	10/10	0.694	0.685	4/10	0.738	0.764	0/10

Table 24: Cross-corpus transfer against HaMI. Train on 80% of the source corpus, score all of the target, 10 seeds.

Model	Transfer	HaMI	PARALLAX	Δ	PARALLAX wins
Llama-3.3-70B	HaluBench → MedHallu	0.884	0.830	-0.054	3/10
Llama-3.3-70B	HaluBench → RAGTruth	0.720	0.668	-0.052	3/10
Llama-3.3-70B	HaluBench → TruthfulQA	0.696	0.695	-0.001	5/10
Llama-3.3-70B	MedHallu → HaluBench	0.855	0.742	-0.113	0/10
Llama-3.3-70B	MedHallu → RAGTruth	0.670	0.709	+0.039	5/10
Llama-3.3-70B	MedHallu → TruthfulQA	0.690	0.755	+0.065	7/10
Llama-3.3-70B	RAGTruth → HaluBench	0.907	0.784	-0.123	0/10
Llama-3.3-70B	RAGTruth → MedHallu	0.959	0.906	-0.053	0/10
Llama-3.3-70B	RAGTruth → TruthfulQA	0.806	0.777	-0.028	1/10
Llama-3.3-70B	TruthfulQA → HaluBench	0.803	0.713	-0.090	1/10
Llama-3.3-70B	TruthfulQA → MedHallu	0.921	0.896	-0.026	3/10
Llama-3.3-70B	TruthfulQA → RAGTruth	0.689	0.644	-0.046	1/10
DeepSeek-R1-32B	HaluBench → MedHallu	0.860	0.808	-0.052	1/10
DeepSeek-R1-32B	HaluBench → RAGTruth	0.697	0.676	-0.021	2/10
DeepSeek-R1-32B	HaluBench → TruthfulQA	0.712	0.719	+0.007	5/10
DeepSeek-R1-32B	MedHallu → HaluBench	0.845	0.753	-0.092	0/10
DeepSeek-R1-32B	MedHallu → RAGTruth	0.665	0.720	+0.055	8/10
DeepSeek-R1-32B	MedHallu → TruthfulQA	0.696	0.729	+0.033	6/10
DeepSeek-R1-32B	RAGTruth → HaluBench	0.850	0.686	-0.164	0/10
DeepSeek-R1-32B	RAGTruth → MedHallu	0.896	0.900	+0.004	5/10
DeepSeek-R1-32B	RAGTruth → TruthfulQA	0.758	0.741	-0.017	3/10
DeepSeek-R1-32B	TruthfulQA → HaluBench	0.832	0.790	-0.042	2/10
DeepSeek-R1-32B	TruthfulQA → MedHallu	0.890	0.845	-0.045	0/10
DeepSeek-R1-32B	TruthfulQA → RAGTruth	0.679	0.665	-0.014	2/10

Table 25: **Sampling-based detectors** (N=10 samples per query) on Llama-3.3-70B (L) and DeepSeek-R1-32B (D), AUROC. PARALLAX uses no samples.

	RAGTruth		TruthfulQA		HaluBench		MedHallu	
Method	L	D	L	D	L	D	L	D
SelfCheckGPT-NLI	0.530	0.526	0.552	0.540	0.722	0.732	0.742	0.761
Semantic entropy	0.477	0.514	0.508	0.465	0.516	0.518	0.496	0.530
SEPs	0.498	0.474	0.506	0.490	0.505	0.569	0.483	0.538
MIND	0.500	0.475	0.510	0.487	0.516	0.576	0.474	0.539
ACT	0.606	0.586	0.516	0.483	0.361	0.269	0.313	0.539
INSIDE	0.546	0.510	0.470	0.484	0.468	0.509	0.513	0.500
HaMI (official)	0.759	0.757	0.917	0.920	0.938	0.946	0.993	0.985
HaMI (official), no sample weights	0.758	0.764	–	–	–	–	–	–
PARALLAX (single pass)	0.746	0.738	0.953	0.943	0.954	0.960	0.994	0.995