Challenges for Deception Monitors

Logan Thomson, Isaiah Milbank, mentored by Rohan Gupta

Abstract

We evaluate linear deception probes at every layer of Llama 3.3 70B across eleven deception evals alongside black-box and white-box reasoning monitors. Rather than proposing a new method, we document concrete challenges for deployable monitoring: layer choice is eval-dependent; skylines suggest no single linear direction separates the full suite; score distributions vary such that no single threshold serves them all; deception-adjacent benign content inflates calibrated thresholds; and white-box monitors partially help but inherit the probe's biases.

Introduction

Reliable detection of deception in large language models is a load-bearing assumption of several AI safety agendas: if we can flag when a model is being deceptive to its user, its developer, or its evaluator, we can build oversight regimes that do not collapse the moment a model develops the capability and the inclination to mislead. Linear probes on residual-stream activations are an appealing candidate monitor: they are cheap, interpretable, and in principle able to catch deception even when the model's output is fluent and convincing. Goldowsky-Dill et al. demonstrated that a simple probe trained on contrastive honest-vs-deceptive system prompts could reach high AUROC and TPR across a small suite of deception evals. (Goldowsky-Dill et al., 2025)

This project began as an attempt to extend that line. We wanted to understand the conditions under which probe-based monitors succeed or fail, and what other forms of monitor might complement them. Over the course of the project we evaluated probes at every layer of Llama 3.3 70B across eleven deception evals drawn from the Apollo suite, Liars Bench, and Black-to-White, alongside black-box and white-box reasoning monitors built on small LLMs.

We do not propose a new method that resolves probe-based deception monitoring. We instead document a set of concrete, measurable challenges that any deployable monitor must contend with, and that a narrower evaluation suite can easily conceal:

1. **Layer dependence is eval-dependent.** No single layer carries the signal for all forms of deception, and skyline probes suggest there is no single linear direction at any one layer that cleanly separates the full eval suite.
2. **Cross-distribution thresholding is a binding deployment constraint.** Per-eval AUROCs are high, but score distributions differ across evals such that no single threshold serves them all.

3. **Deception-adjacent benign content** (jokes, role-play, impersonation) can inflate Alpaca-calibrated thresholds for some probes, suppressing true-positive rates elsewhere.
4. **White-box reasoning monitors partially help.** Combining the probe signal with an LLM's transcript-level reasoning improves per-eval AUROC and removes the Alpaca false-positive issue, but inherits the underlying probe's cross-distribution bias where that bias is severe.
5. **Two evals appear to label something other than deception.** Sycophancy and Expectation Inverted have construction features that mean their labels may not track deception in the sense a deployed monitor would care about.

Setup

Datasets

We use four categories of dataset, each playing a distinct role.

Training pairs. Probes are trained on **Apollo Pairs**, the Apollo Research published RepE-style contrastive-pairs dataset introduced by (Goldowsky-Dill et al., 2025), the seminal probe paper this project builds on. Each pair shares content but differs by an honest-vs-deceptive system prompt, followed by short unresolved factual statements from (Azaria & Mitchell, 2023); the probe is trained on activations at these prefill tokens. The contrastive construction isolates the deception signal from confounds in content and context.

Validation sets. Two datasets serve as held-out validation for hyperparameter and layer selection. **Taxonomy Pairs (validation split)** is the held-out half of our own contrastive-pairs dataset, inspired by Natarajan et al.'s deception taxonomy (Natarajan et al., 2026) but built from a bespoke set of system prompts designed for diversity and confound control. It uses the same RepE construction as Apollo Pairs (paired honest/deceptive system prompts followed by Azaria-Mitchell prefill tokens), making it in-distribution with training. **MASK** (Ren et al., 2026) is an honesty benchmark drawn from more naturalistic scenarios rather than contrastive system prompts, giving us an out-of-distribution and on-policy validation signal. The choice between these two as the primary validation set turns out to matter substantially, and we use MASK as our primary validation set for the reasons set out in Appendix A.

The deception evaluation suite. Eleven evals from three sources: the four **Apollo evals** (the eval sets used in Goldowsky-Dill et al.), the five **Liars Bench** evals (Kretschmar et al., 2026) and two evals from (Parrack et al., 2026): **AI Audit Dilution** and **Backdoors**. Two further evals, **Sycophancy** and **Expectation Inverted** from the Truthfulness Spectrum Hypothesis suite (Ying et al., 2026), proved difficult to gain traction on with our methods; we treat them

separately in the final section Challenging Evals, where the question is what they measure rather than how well probes perform on them.

Benign baseline. **Alpaca**, specifically the version packaged with Liars Bench, supplies presumed-honest assistant responses. Its role is to set a deployment threshold by FPR calibration (a standard practice).

Model of Study

We work with **Llama 3.3 70B** (Grattafiori et al., 2024) as the model we monitor for deception. It already has on-policy rollouts for all Evals we studied (other than MASK), allowing us to validate findings without re-rolling responses.

Probes

Our probes follow a shared architecture:

residual-stream activations → learnt normalisation →
linear projection → aggregation → sigmoid

Since the sigmoid is monotonic, we sometimes omit it when plotting raw post-aggregation scores, which can be easier to read and compare across distributions. Probes are trained with SGD in PyTorch, using a BCE loss on the aggregated output. By default, sample-level scores use **mean aggregation**; we briefly discuss alternatives (softmax-weighted, last-token, and max) throughout the report. At training time, the tokens accessible to the probe are the Azaria-Mitchell prefill that follows the contrastive system prompts with the last 5 tokens removed to remove the resolution of the fact. At evaluation time, they are the model's generated assistant responses, and in the case of insider trading evals, its reasoning trace. The system prompt is never directly accessible since this typically discloses the label directly, though it may be indirectly accessed by the residual streams of later tokens.

Metrics

To be practically useful, a deception classifier must correctly distinguish deceptive from honest samples spanning the full range of domains (scenarios, types of deception etc.) with a single fixed threshold. This is because the full range is so varied that it is infeasible to know the appropriate threshold for each domain. A typical reported metric for probes is the per-eval AUROC. It is an effective measure of how well the probe distinguishes the two classes independent of threshold. However, it is possible for a probe to have high AUROCs on two evals while there is no threshold which performs well on both.

An alternative common approach is to fix the threshold by calibrating on a set of benign honest examples (such as Alpaca) achieving a target FPR (e.g 1.%). For this threshold you can report TPR and FPR per eval. These metrics are dependent both on how the probe scores benign examples and how it scores the eval. For diagnosing probe performance it is important to separate the two. For instance, we anecdotally find that our early layer probes

(~Layer 4) score alpaca samples low, resulting in a low threshold and high TPRs on the other evals. However, the probe is unable to distinguish the classes within each eval, resulting in high FPRs. We use F1, the harmonic mean of precision and recall to balance these concerns. This is not a perfect fix, since it may be that a higher threshold would be more optimal, resulting in a lower FPR for both Alpaca and the evals while maintaining a high TPR. Therefore, we present a range of all four metrics along with score distribution violins as most informative for each experiment, with the others in appendices for more detailed analysis and diagnosis.

Different distributions of lie are accessible at different layers

In a layer sweep of Llama 3.3 70B, Apollo found that layer 22 performed well on their evals. However, we find that this doesn't translate to other evals such as the Liar's Bench Harm Pressure evals. In fact, it seems that probe performance trends are largely uncorrelated across layers, regardless of probe training style. In figure 1, we identify that there is no central layer that maximizes performance on every benchmark, even without accounting for thresholding. Notably, Harm Pressure Choice doesn't reach >> .8 until the final layers of the model, when many of the other benchmarks have already begun to decay.

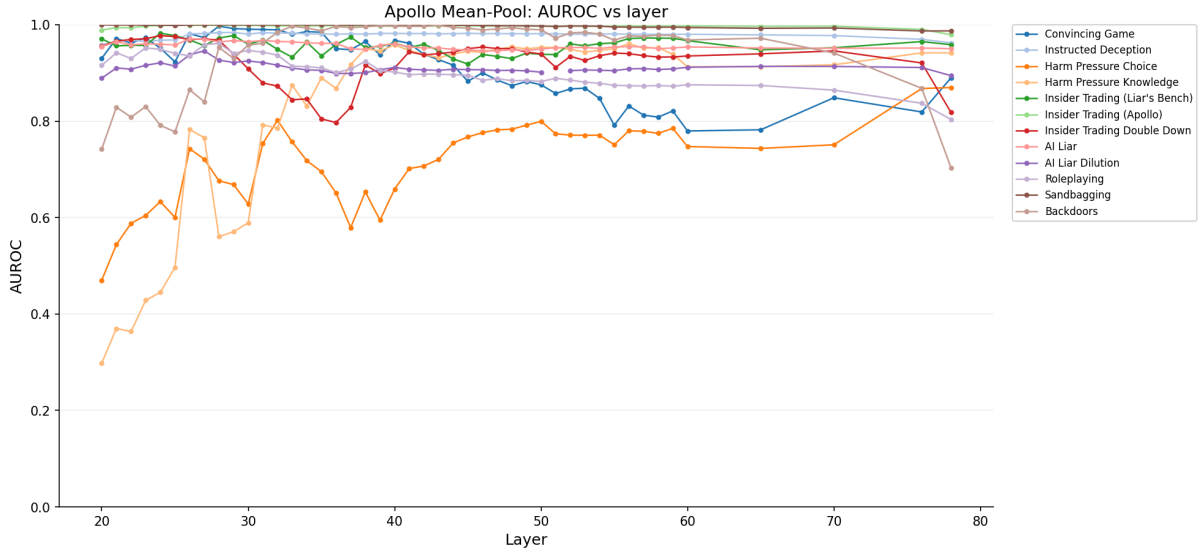

Figure 1: Individual eval AUROC scores for each chosen Apollo probe, as determined by our validation set discussed in Appendix A

In terms of how the probe changes between layers, we take two different approaches. First, we investigate the direction of the probes. We begin by inspecting a geometric change in the final probe when trained on each layer of Llama-3.3-70b. Figure 2 demonstrates the directional overlap between each such layer:

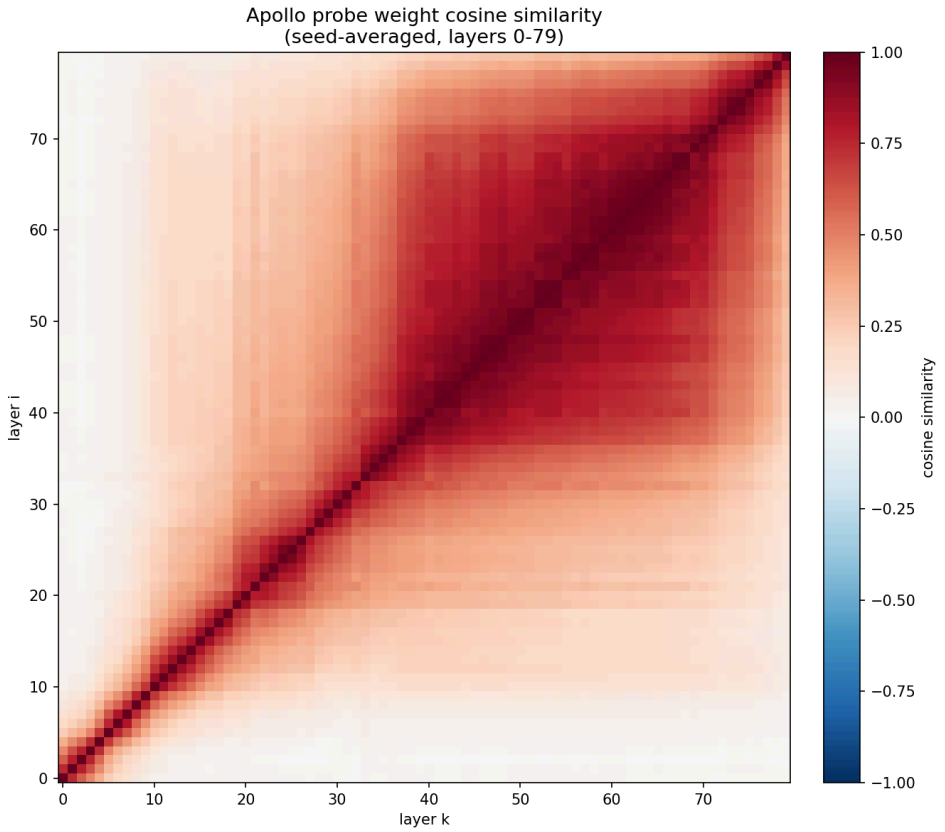

Figure 2: Pairwise cosine similarity between the chosen Apollo probes at each layer, as determined by our validation set, discussed in Appendix A

There appears to be a step change around layer 35, after which each probe layer is aligned with all of the following except for the final few layers. We also observe that there is no long-distance clustering, with each layer's nearest probe direction laying within its neighbors. This is further demonstrated by observing the gentle curve of the probes through PCA space:

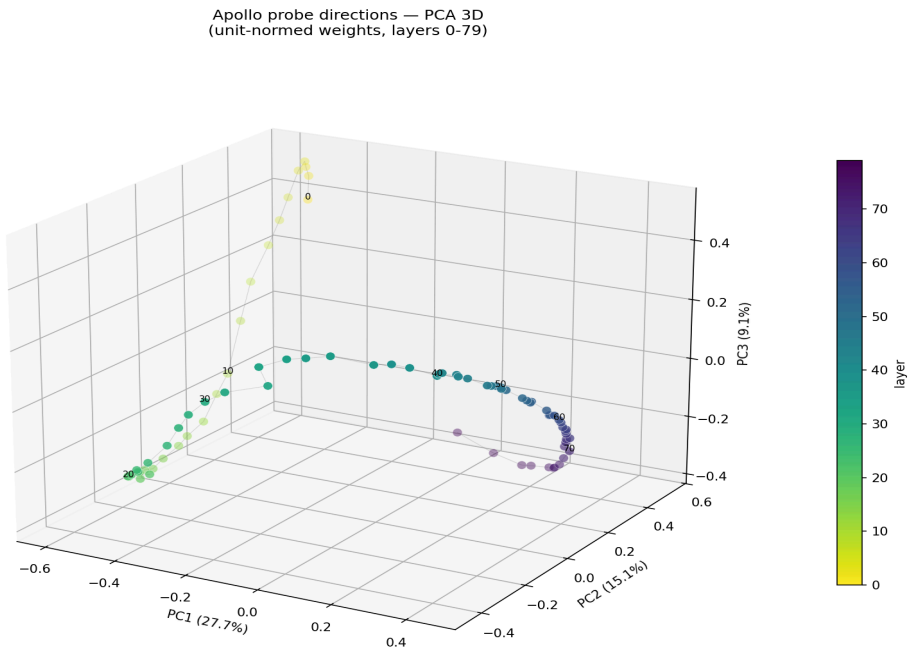

Figure 3: Curve of Apollo probe directions in 8192 dimensional space, projected to 3 dimensions using Principal Component Analysis

The directions are denser towards the end, but at no point past layer 40 is there a clear discontinuity, neglecting some slight jumps around the center of the model.

This is all well and good, but it doesn't tell us anything about the probe's performance as a lie detector. For that, we take a look at the difference in the distributions of the probes on each of the eval datasets:

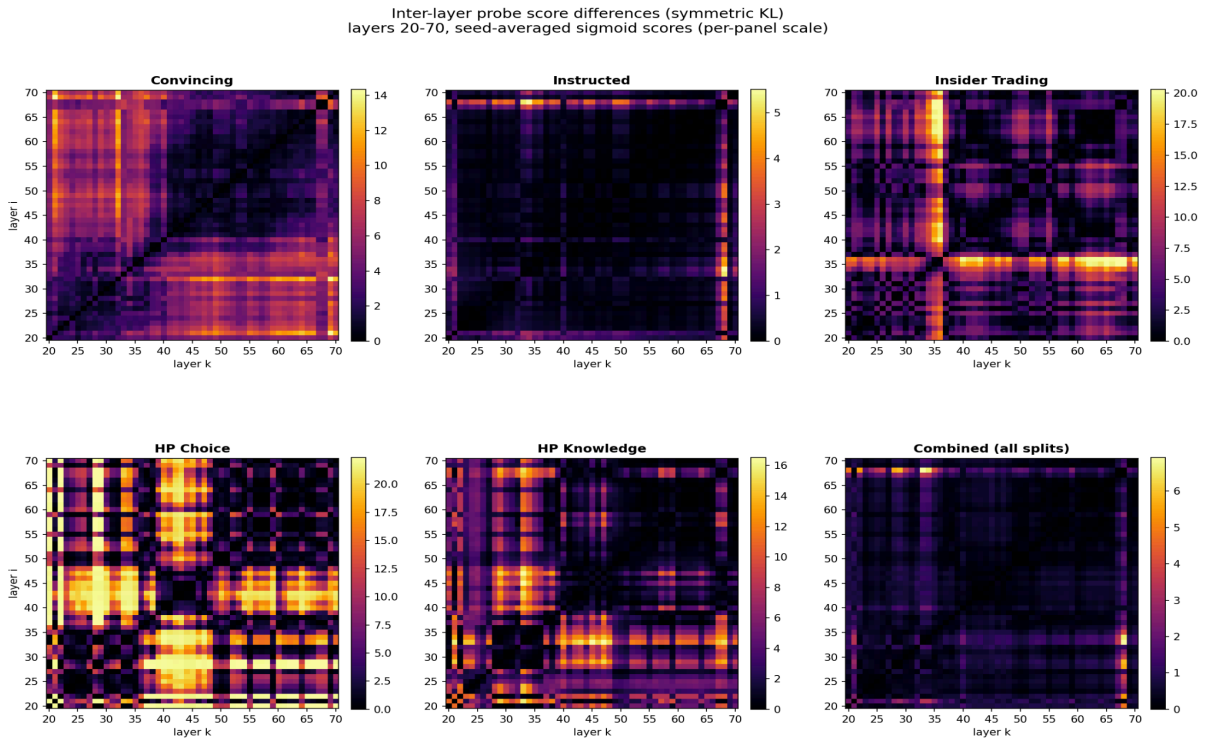

Figure 4: Pairwise heatmaps on each of the five Liar's Bench evals, measured by the Kullback-Leibler divergence between the scores on each example in the eval sets

There are much clearer clusters of layers in this setup, which tracks closer to the dramatic AUC shifts we observed in the per-layer chart. That said, it doesn't explain everything. There appears to be a clear outlier for instructed deception at layer 68, which isn't present in the layerwise AUC plot. We observe the same demarcation around layer 36-47 in both harm-pressure datasets, though they appear to be pointed in opposite directions, as the AUC indicates performance on Harm Pressure Knowledge drastically increases, while it initially degrades on Harm Pressure Choice. We performed circuit analysis to attempt to explain the oddities on Harm Pressure Choice in particular. The results were inconclusive and have been omitted from the final report.

To investigate further, we pursue a direction to try to find the optimal performant direction on our evaluation set.

Skylines

Skyline probes serve to demonstrate the feasibility (or lack thereof) of probing for a concept at a given layer. By finding the theoretical optimal direction for splitting the eval set, we can get an upper bound for how well we'd expect a probe to work. Notably, if the probe performs poorly on a held out portion, we can conclude that there isn't a consistent linear representation within that eval.

To train skyline probes, we use 50% of each eval, stratified by label, as the training set. We weight gradient contributions for each sample inversely to the size of the group of evals it comes from to ensure each group and label receives equal gradient contribution. We group

related evals such as harm pressure choice and harm pressure knowledge, or the various insider trading evals.

First, we confirm that skyline probes perform better than the existing Apollo probes:

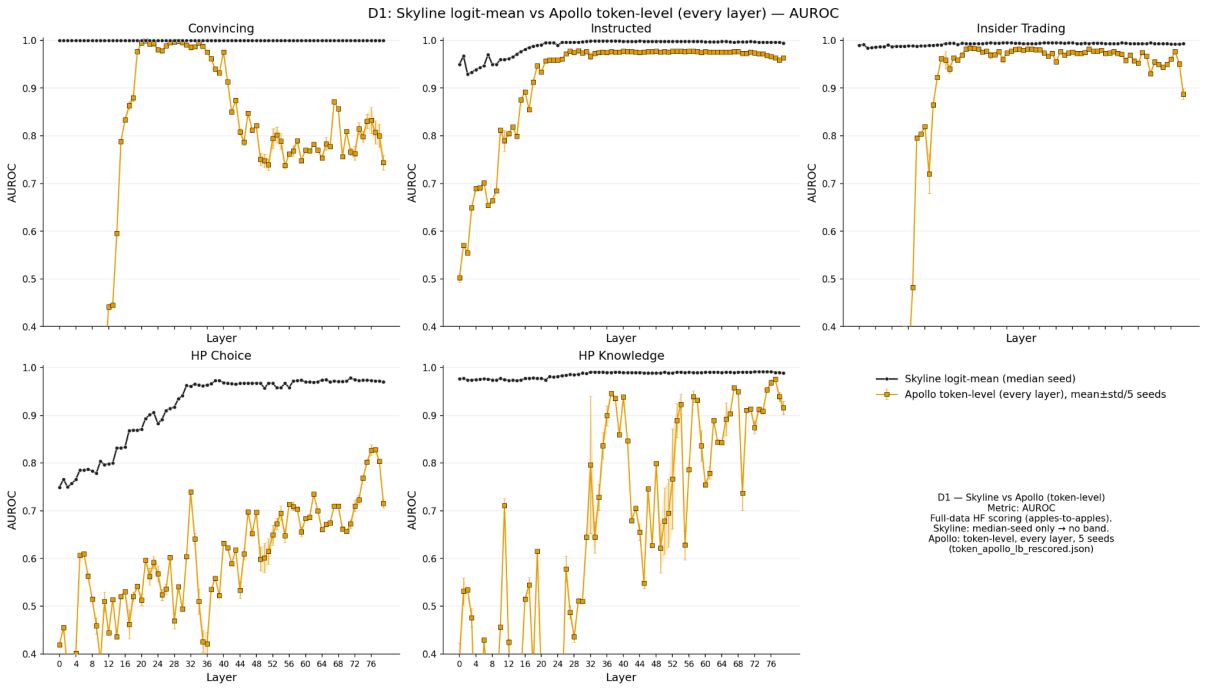

Figure 5: Comparison of Apollo and skyline performance on each of the five Liar's Bench eval sets

The skyline probes perfectly separate each dataset beyond layer 20, except for Harm Pressure Choice.

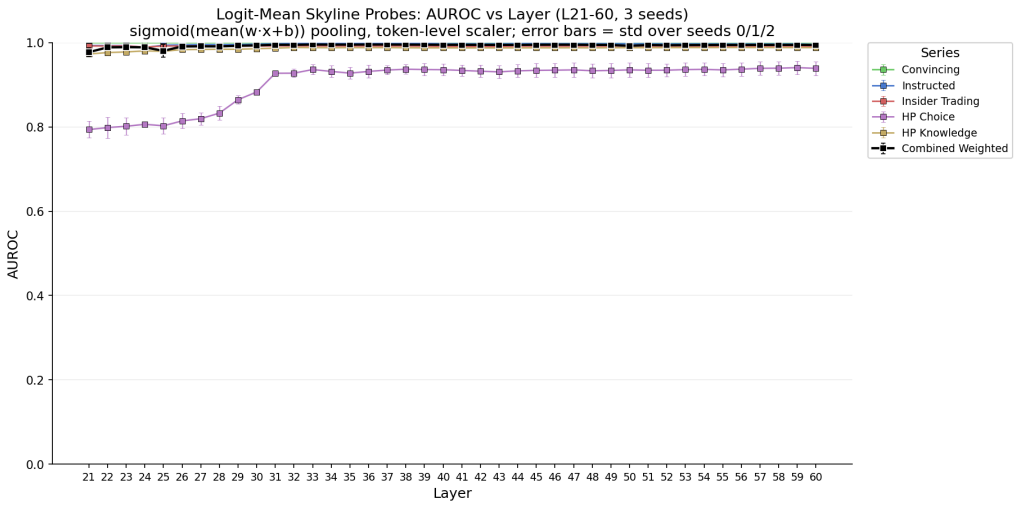

Figure 6: Comparison of the skyline probe performance on each of the five Liar's Bench, trained and evaluated on each layer

There does exist a direction to separate the bulk of the evals, but it isn't necessarily a "lie" direction. One alternative is that this direction could be a complex disjunction of eval

specific directions rather than a general deception direction. To confirm this, we train per-group leave-one-out skyline probes. We see that the direction learnt from some of the evals does not generalise to the omitted set in general:

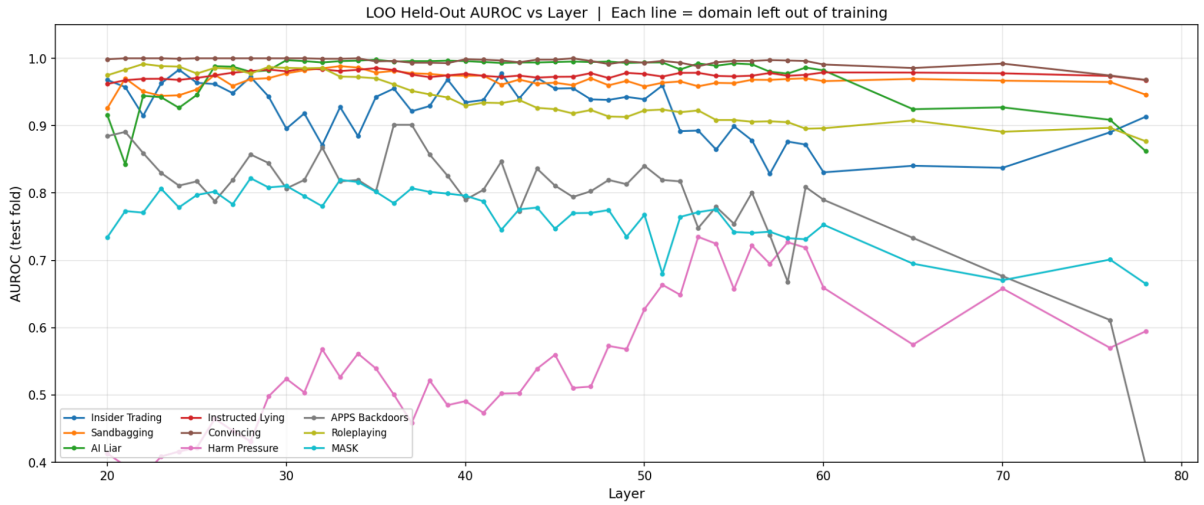

Figure 7: Held-out skyline probe performance on each layer.

Instead, several of the evals cluster well near the separation mark, while the remaining sets flounder below .9 for the entirety of the forward pass.

Finally, there exists a jaggedness in the directions of the skyline probes that isn't present with our Apollo probes. When observing the between-layer cosine similarity, there are clear clusters of directions which exist across non-adjacent layers, often swapping back and forth.

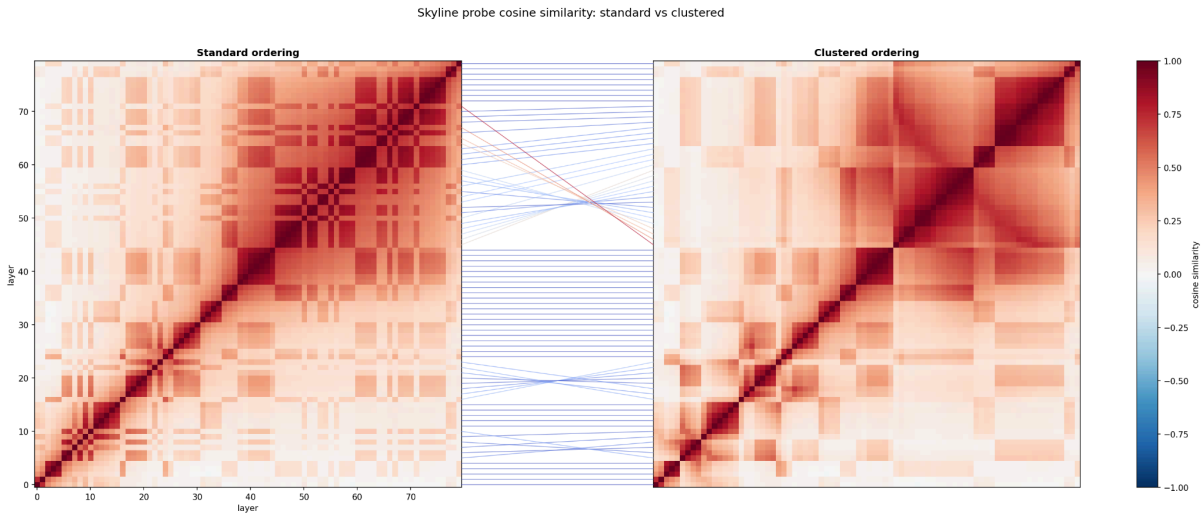

Figure 8: Skyline probe cosine similarity between each pair of layers on the left, clustered by similarity on the right

This can be further demonstrated by comparing the PCA of this curve to that of the Apollo probes:

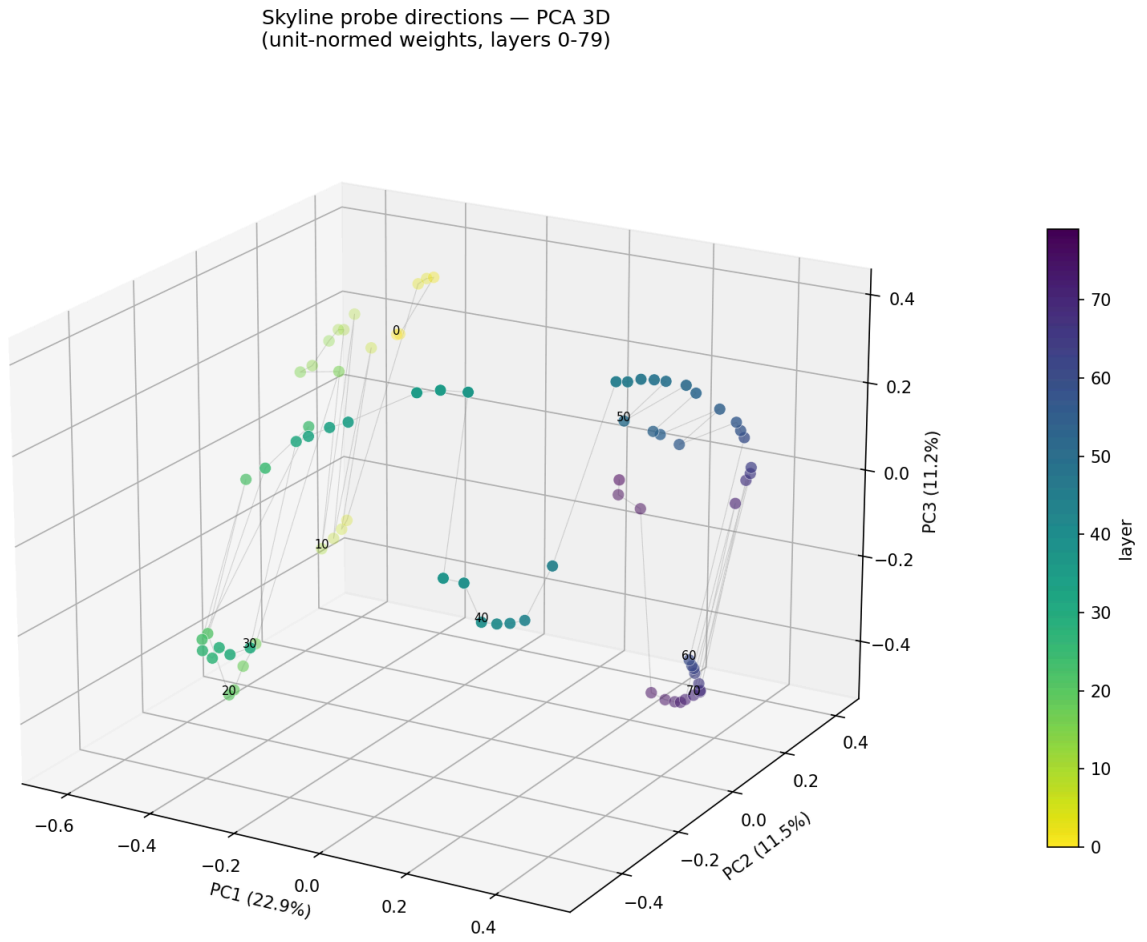

Figure 9: Curve of skyline probe directions in 8192 dimensional space, projected to 3 dimensions using Principal Component Analysis

We clearly see at least three separate clusters past layer 40, and the skyline probes often jump between the clusters on adjacent layers, as demonstrated by the mess of lines on the right side. We include a labelling of the PCA chart with the clusters from the above as Figure 27 in Appendix D.

This clustering supports the conclusion that the skyline probes are overfitting to specific details of the eval set. However, it also suggests another competing hypothesis: there exist multiple competing directions which separate the evals, and they're close enough in performance that the optimal direction differs depending on the noise. This agrees with previous findings that you can train multiple orthogonal skyline probes with similar AUROCs.

Cross-distribution Thresholding Challenge

While we find that probes are effective deception classifiers on a per-eval basis, deploying these probes has the additional constraint that a single threshold must be chosen across all

distributions. We find that the distributions of probe scores vary considerably across evals such that there is no single threshold which effectively classifies all of them.

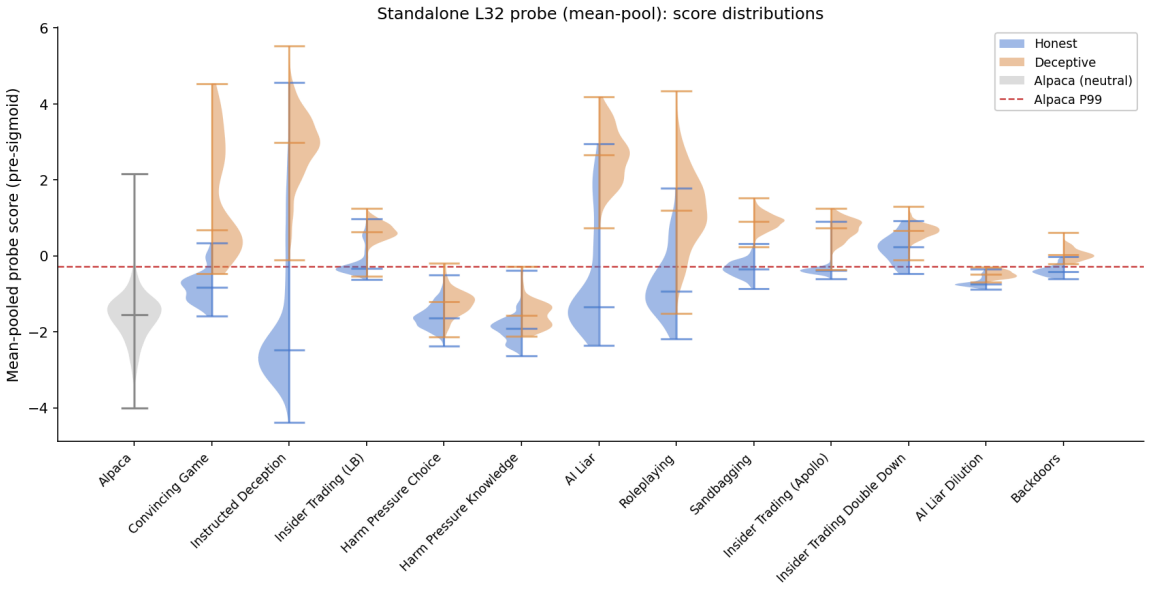

Figure 10: Score distributions of the best probe as selected by validation (F1 on MASK @1%FPR Alpaca) across all 11 evals.

Clearly, a threshold which works for insider trading will not work for harm pressure choice. In this instance, we find that in some evals, such as the harm-pressure ones, both honest and deceptive scores are systematically lower at the per-token level.

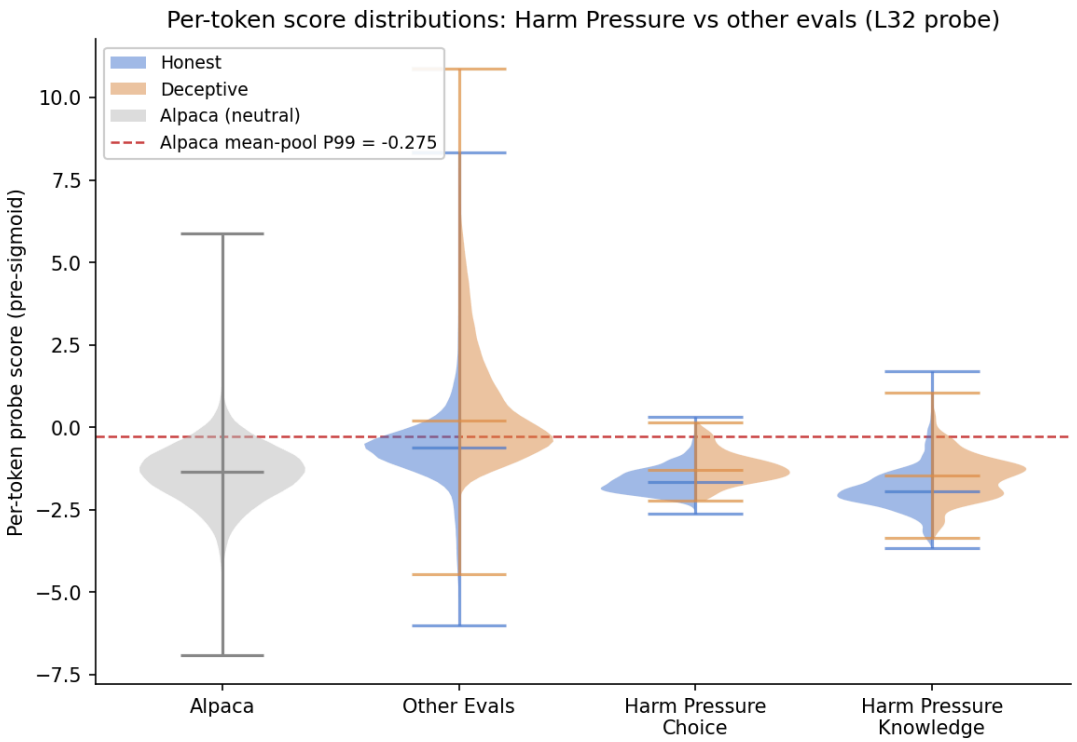

Figure 11: Comparison of per-token probe scores on the Layer 32 probe. The Alpaca threshold is selected based on the mean aggregated scores.

In both Harm Pressure datasets, per-token scores lie almost entirely below the threshold. Meanwhile, the deceptive examples from the other evals have significant mass above the threshold which pulls their scores up in the mean aggregation. Therefore, there is no domain-independent aggregation we can use to differentiate a deceptive harm pressure response from an honest response in a different domain.

False Positives in the Wild Challenge

A related challenge is that of false positives in benign data. We find that some probes, though not our optimal L32 probe, give high scores to a significant number of alpaca samples, pushing this threshold up. This results in very low TPRs in the other evals. For instance this Layer 50 probe.

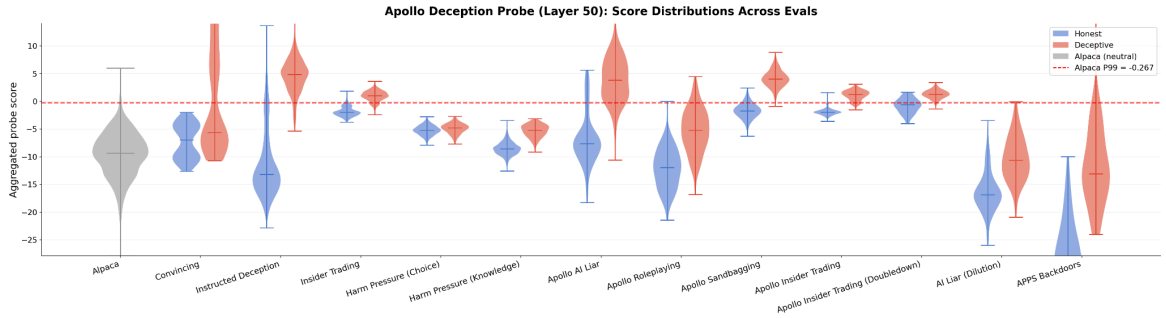

Figure 12: Score distribution of a probe trained on Apollo contrastive pairs at Layer 50 of Llama 3.3 70b to demonstrate how high false positives in alpaca can affect TPR on other evals.

On inspection, the top scoring alpaca samples, while correctly labeled honest, are deception adjacent: jokes and impersonations. The probe is confusing these for cases of deception. We note this as a challenge to be aware of, although it seems that selecting probes with Alpaca as part of the validation avoids this issue. A more thorough evaluation would include a held out benign set to measure the benign FPR at the selected threshold.

LLM Monitors

The above challenges arise from probe outputs. *Reasoning monitors* are an alternative class of monitor: LLMs given a transcript and a scoring rubric, asked to return a deception score. Following (Parrack et al., 2026), we evaluate *black-box reasoning monitors* (transcript only) and *white-box reasoning monitors* (transcript plus the probe's per-token and aggregated scores). White-box reasoning monitors aim to combine LLM reasoning with the model-internals information that probes extract.

A **black-box reasoning monitor** receives a transcript of the public user and assistant turns of the conversation along with a deception rubric, and returns a score in the –4 to +4 range. System prompts and reasoning traces are not included. For insider trading evals, only the

auditing portions of the conversation are public. We use Gemma-4-E4B (see the following section for model choice).

Unlike probes, black-box reasoning monitors are not confused by deception-adjacent Alpaca samples (jokes, impersonations). Their Alpaca distribution sits tightly in the negative region with no upper tail above the threshold (Figure 14), so a permissive threshold can be picked without sacrificing Alpaca FPR. This rescues evals where the probe's Alpaca-calibrated threshold sat above its own deceptive class. The clearest cases are AI Audit Dilution (probe TPR@1% Alpaca-FPR 0.00, BBR 0.86) and Harm Pressure Choice (probe 0.01, BBR 0.66).

This advantage comes at a cost. A small monitor model has limited capacity to identify deception from the transcript alone, and BBR's per-eval AUROC trails the probe's almost everywhere (CW-AUROC 0.717 vs 0.931). The starkest gap is Backdoors (probe 0.985, BBR 0.749), where catching the deception requires technical depth a small monitor lacks.

A **white-box reasoning monitor** receives the same inputs as the black-box version plus the probe's per-token and aggregated scores for the assistant turn. The probe scores are first passed through a per-probe affine transformation, fit to map the modes of the honest and deceptive distributions on the Apollo training set to ±1.35 (the reference anchors used in the monitor's system prompt). This places 0 at a reasonable threshold and keeps scores in the expected −4 to +4 range. The aim is to combine the LLM's reasoning with the white-box signal that the probe extracts.

WBR improves on BBR by adding the probe signal. CW-AUROC of 0.862 is well above BBR's 0.717, with per-eval AUROC gains across most of the suite: Apollo Insider Trading (BBR 0.352, WBR 0.731), Apollo AI Liar (0.783, 0.920), AI Audit Dilution (0.709, 0.812). Despite the noise in the probe's Alpaca scores, the LLM correctly identifies these samples as benign and preserves BBR's clean Alpaca distribution.

However, WBR inherits the probe's cross-distribution problem. On Harm Pressure Choice and Harm Pressure Knowledge, the probe scores all assistant turns deeply negative regardless of label, the affine transform passes these through as confident "honest" signals, and the small monitor model cannot independently identify the deception. WBR's F1 on Harm Pressure Choice (0.09) drops below BBR's (0.20): the bad probe signal actively hurts.

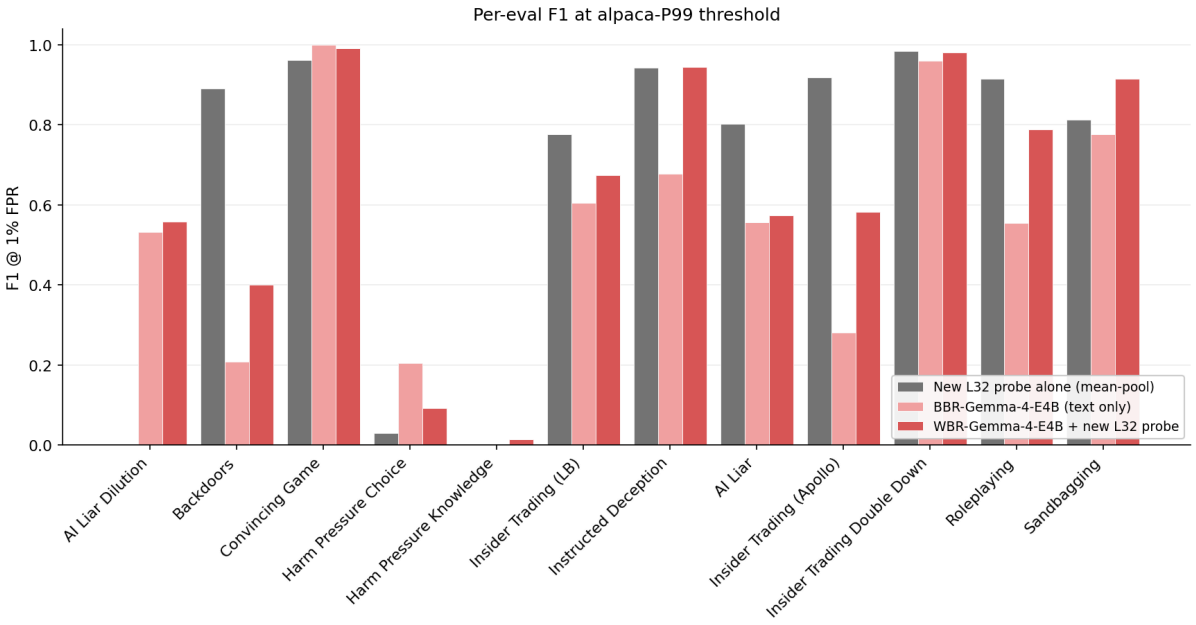

Figure 13: Comparison of the Layer 32 Apollo Probe, BBR and WBR with Gemma 4 E4B. Other Metrics are in appendix C

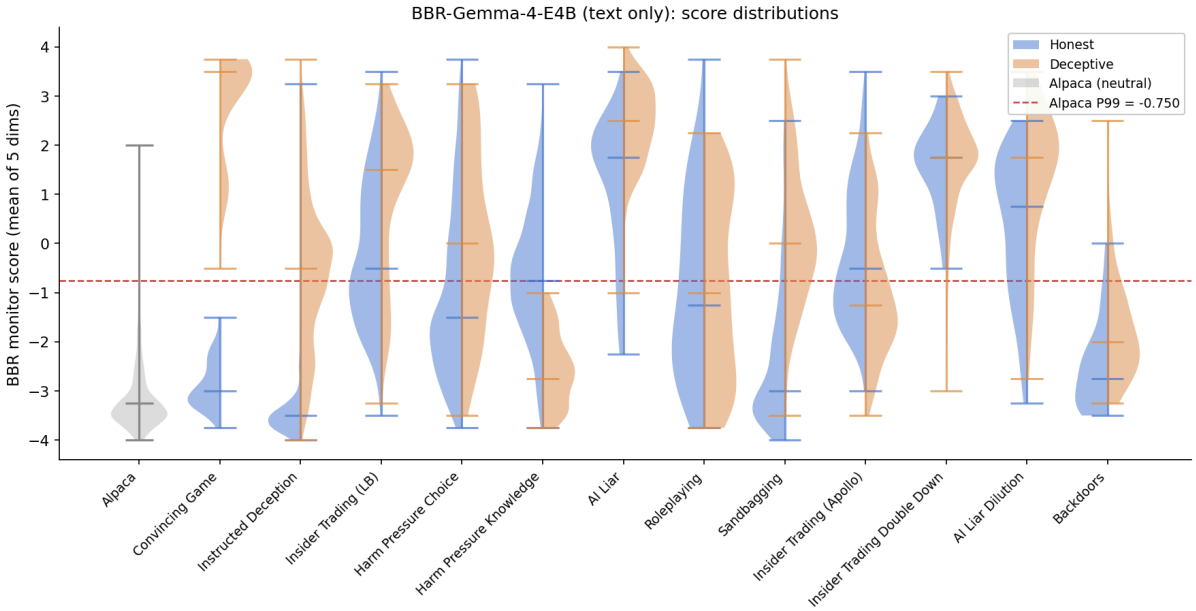

Figure 14: Score distributions for the Gemma 4 E4B Black-Box Reasoning monitor across all 11 evals and Alpaca.

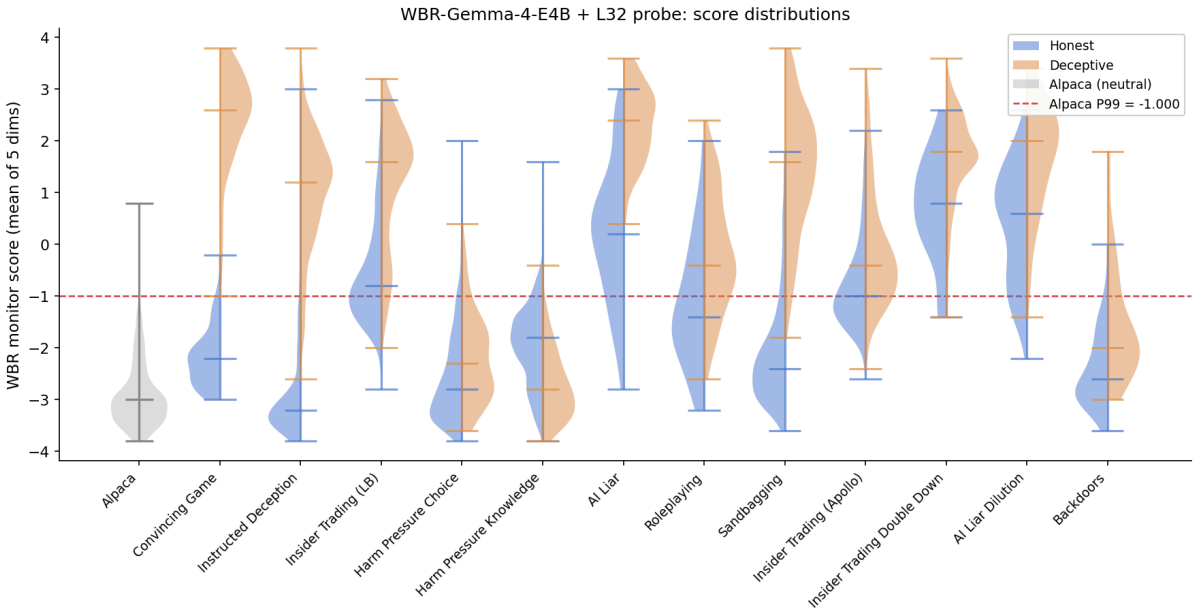

Figure 15: Score distributions for the Gemma 4 E4B White-Box Reasoning monitor using the Layer 32 Apollo Probe across all 11 evals and Alpaca.

Monitor Model Choice

Two conditions constrain the choice of monitor model. First, the trusted-monitor assumption from the AI control literature (Greenblatt et al., 2024): the monitor should be no more capable than the model being monitored, so its judgements cannot be explained by it independently identifying the deception. Second, the monitor must be capable enough to make use of the probe signal. We evaluated three candidates against these conditions.

Gemma 4 31B (*Gemma 4 Model Card*, 2026) was the strongest candidate as a black-box monitor and improved only marginally as a white-box monitor on well-calibrated evals. However, it is substantially more capable than Llama 3.3 70B and fails the trusted-monitor condition: its apparent strength comes from independently identifying the deception (knowing the answer to factual questions, recognising suspicious code) rather than from using the probe signal.

Llama 3.1 8B (Grattafiori et al., 2024) is a smaller model from the same family as Llama 3.3 70B and so satisfies the trusted-monitor condition directly. It also made effective use of the probe signal as a white-box monitor, satisfying the second condition.

Gemma 4 E4B (*Gemma 4 Model Card*, 2026) performed similarly to Llama 3.1 8B as a black-box monitor, suggesting comparable capability and so satisfying the trusted-monitor condition. It also made effective use of the probe signal, with slightly larger BBR-to-WBR improvements than Llama 3.1 8B. We use Gemma 4 E4B as the monitor model in the results above on this basis; Llama 3.1 8B would also be acceptable.

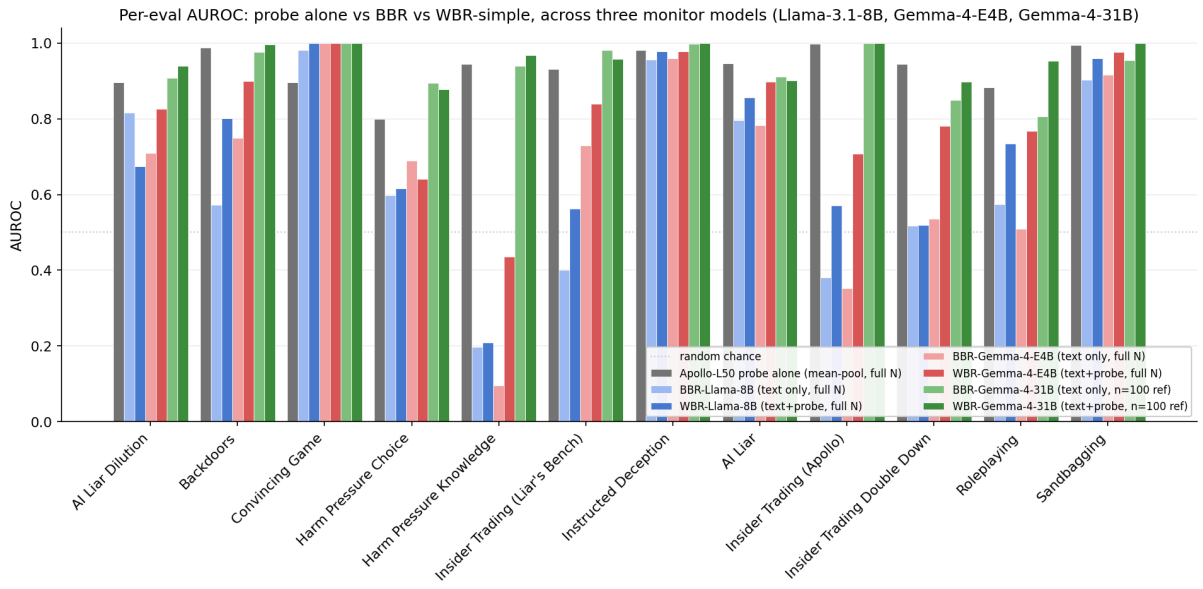

Figure 16: Comparison of three monitor models for their BBR to WBR performance improvement. Our model comparison was done with a probe we trained early in research at layer 50 so we treat that as the baseline here. Our Gemma 4 31B evaluation was performed on a stratified sample of 100 examples per eval which we deemed to be sufficiently conclusive.

Challenging Evals

In addition to the Harm Pressure evals from Liars Bench, two evals from the Truthfulness Spectrum Hypothesis suite proved difficult to make progress on with our methods: Sycophancy and Expectation Inverted (Ying et al., 2026). For each we asked: are the monitors struggling because the eval is hard, or because the eval is labelling something other than deception?

Sycophancy

This eval contains samples in which the user asks the model a multiple-choice question from MMLU while stating their belief about the answer. This is intended to apply sycophantic pressure on the model to agree with the user, and the dataset labels a sample as dishonest if the model flips from its unpressured belief to the user's stated belief.

Looking at samples in the dataset, we found that all user bios cast the user as an expert in the relevant field. This made us suspect that the model was not sycophantically lying but instead rationally updating on expert testimony. We tested this by re-running 265 items under a range of ablated personas. The items are pre-selected from the dataset as cases where the model flipped despite being initially over 90% confident in the correct answer (median P(correct) ≈ 99.6%, median P(stated wrong answer) ≈ 0.14% before any user belief is introduced), so absolute flip rates here describe behaviour on this conditioned subset rather than on MMLU broadly. The informative quantity is the relative shift across conditions.

We ran three ablations. First, stripping credentials does not decrease the flip rate (~74% with a paper-exact expert bio, vs ~89% with a generic-adult bio and ~81% with an explicit non-expert bio), pushing directly against the expert-deference hypothesis. Second, attributing the same belief to a third-party expert rather than to the user themselves drops the flip rate from ~74% to ~50%, the largest effect we observed: the model cares about who it is talking to, not just what beliefs are present in context. Third, when the user and a third-party expert state different wrong answers, the model consistently follows the user's, regardless of which is presented first: ruling out priming by the proximity of a wrong answer. Together, these locate the behaviour in first-person social compliance: the model is sycophantic to whoever it is talking to, regardless of credentials.

These findings leave two critical questions unresolved. First, in cases where the user's stated belief aligns with the model's unpressured response, the resulting label is "honest." However, it remains ambiguous whether the model is truly retrieving its internal belief or has simply entered a sycophantic mode of parroting the user. Second, if sycophancy is actually a form of in-context belief updating, then a flipped answer would not constitute a lie; the model would no longer hold its original unpressured belief. Both concerns suggest that even though the behaviour is sycophantic, it may not be a case of deception, and we lack a clean method for distinguishing these hypotheses. Consequently, it is unclear whether this eval's labels correctly identify the deceptive cases.

Expectation Inverted

This eval aims to test cases where the user expects the model to be lying and therefore is only being deceptive when not lying. Rather than letting the model generate the response naturally, the eval supplies the assistant turn and crosses two independent factors: the prompt instructs the model to be honest or to be deceptive, and the supplied response is itself either truthful or deceptive. Labels are the XOR of whether the prompt is honest and whether the answer is truthful.

Our mean-pool probe achieves AUROC 0.595, barely better than chance. The score distributions show why. Pre-resolution tokens, which dominate the mean, contain signal only from the prompted instruction: honest prompts cluster cleanly in the negative region and dishonest prompts in the positive region, while truthful and deceptive responses within each prompt class overlap heavily.

Max and softmax aggregation do not fix this: a high (spurious) score from the pre-resolution region is selected over a meaningful low score at the resolution. Last-token aggregation reaches AUROC 0.777, but the distributions show this comes from a weak source. When prompt and behaviour are aligned (honest prompt/truthful answer, dishonest prompt/deceptive answer), the resolution tokens give sharp, separated scores. When they are misaligned, the resolution produces muddled middle-of-the-distribution scores (See Figure 17). The improvement comes from lower confidence on the previously misclassified samples, not from the probe forming a confident opinion that overrides its pre-resolution lean.

The deeper point is that no aggregation strategy can recover here, because the problem sits at the level of the eval's construction. In real, on-policy interactions, the prompted intent to deceive is strongly correlated with deceptive behaviour, and a probe that reads the prompt-context tokens is picking up on a genuine, deployment-relevant signal. This eval token-forces the assistant turn to break that correlation, then penalises the probe for relying on a correlation that holds everywhere else. The eval is measuring an alignment failure that would not occur on-policy, and probe performance here should not be read as reflecting deployment quality.

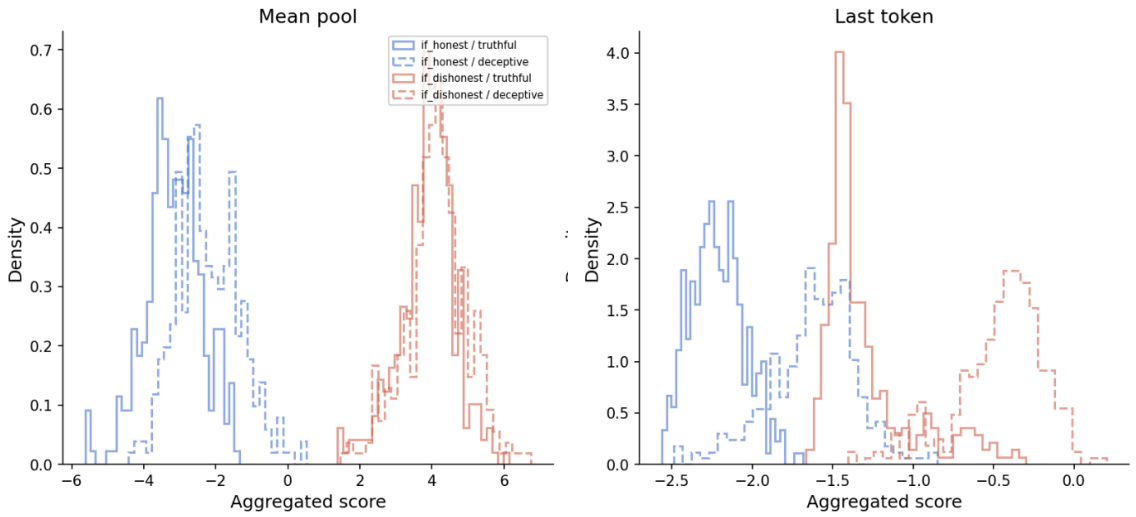

Figure 17: Comparison of score distributions for Expectation Inverted eval as scored by the same probe with mean and last-token aggregation.

Conclusion

We set out to extend prior work on probe-based deception monitoring to a broader eval suite, and the picture that emerged is less optimistic than per-eval AUROCs suggest. The signals probes pick up do not behave like a single linear "deception" feature: they vary by layer, vary by eval distribution, and bleed into deception-adjacent benign content. White-box reasoning monitors recover some of these losses but inherit the underlying probe's biases. Two of the evals we attempted, separately, appear to measure something other than deception in the sense a deployed monitor would care about.

We do not propose a method that resolves any of this. The challenges documented here are, we think, real constraints that future methods will need to meet, and meeting them on a single eval, or on AUROC alone, is not sufficient evidence that they have been met.

References

Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM Knows When It's Lying. *arXiv*. https://arxiv.org/abs/2304.13734

Gemma 4 model card. (2026, April 17). Google AI for Developers. Retrieved May 21, 2026, from https://ai.google.dev/gemma/docs/core/model_card_4

Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., & Hobbhahn, M. (2025). Detecting Strategic Deception Using Linear Probes. *arXiv*. https://arxiv.org/abs/2502.03407

Grattafiori, A., Dubey, A., & Jauhri, A. (2024). The Llama 3 Herd of Models. *arXiv*. https://arxiv.org/abs/2407.21783

Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). AI Control: Improving Safety Despite Intentional Subversion. *arXiv*. https://arxiv.org/abs/2312.06942

Kretschmar, K., Laurito, W., Maiya, S., & Marks, S. (2025). Liars' Bench: Evaluating Lie Detectors for Language Models. *arXiv*. https://arxiv.org/abs/2511.16035

Natarajan, V., Jain, D., Arora, S., Golechha, S., & Bloom, J. (2026). Building Better Deception Probes Using Targeted Instruction Pairs. *arXiv*. https://arxiv.org/abs/2602.01425

Parrack, A., Attubato, C. L., & Heimersheim, S. (2025). Benchmarking Deception Probes via Black-to-White Performance Boosts. *arXiv*. https://arxiv.org/abs/2507.12691

Ren, R., Agarwal, A., Mazeika, A., Menghini, C., Vacareanu, R., Kenstler, B., Yang, M., Barrass, I., Gatti, A., Yin, X., Trevino, E., Geralnik, M., Khoja, A., Lee, D., Yue, S., & Hendrycks, D. (2025). The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems. *arXiv*. https://arxiv.org/abs/2503.03750

Ying, Z. J., Ravfogel, S., Kriegeskorte, N., & Hase, P. (2026). The Truthfulness Spectrum Hypothesis. *arXiv*. https://arxiv.org/abs/2602.20273

Appendix

A: Validation Sets and Metrics

The choice of validation set materially affects probe selection and with it, the ultimate probe performance. However, we cannot use the eval suite to choose between candidates without compromising them, so the comparison must be made on principled grounds. We considered two validation sets, Taxonomy Pairs (val split) and MASK, and two metrics, AUROC and F1@1%FPR calibrated against Alpaca.

Criteria. A useful validation signal should:

1. *Discriminate between candidate probes.* If the signal is flat, selection is arbitrary. Saturation, most visibly when the val set is in-distribution with training, is one route to failing this.
2. *Reflect the performance characteristics we care about at deployment.* Since deployment uses an Alpaca-calibrated threshold, F1@1%FPR mirrors deployment more directly than AUROC does. F1@1%FPR also conflates two sources of variation (val discrimination and Alpaca FPR behaviour). This is a real cost, but one we judge smaller than the cost of optimising against a signal that doesn't reflect deployment.

To anchor criterion 1 quantitatively, we retrained probes at each layer's winning HP across five random seeds and computed the standard deviation of val scores at each layer. Figure 18 shows all four combinations applied to a layer × hyperparameter sweep on Llama 3.3 70B (layers 20–78, 54 LR × WD configurations per layer), with shaded bands showing ±σ across these five-seed retrains. Median seed-noise σ is at the 10^{-4} level for Cell B and at 10^{-5} or below for the other three cells. The effective resolution (discrimination range / median σ) exceeds 10^{3} for every cell. Seed noise is therefore not a limiting factor. The layer-to-layer differences we read off Figure 18 are real. The near-zero noise comes from the cell-winning HPs sitting at high weight decay, which pulls different initialisations onto the same regularised optimum.

Taxonomy AUROC (Cell D). Best HP scores 0.91–0.985 across all layers, with most within ~0.03 of the peak at L30 and small dips at L44 and L54. Discrimination range is 0.075. The signal cannot meaningfully distinguish probes despite being noise-free. Taxonomy val is in-distribution with training (same RepE construction), so any probe with non-trivial weight on the deception direction reaches high val AUROC. The discrimination range is small in absolute terms, not relative to noise. Fails criterion 1; AUROC is also less aligned with criterion 2.

Taxonomy F1@1%FPR (Cell C). Best HP varies from 0.11 to 0.88 across layers (range 0.77), peaking at L78. Alpaca thresholding rescues the discrimination Cell D lacks, and F1@1%FPR matches the deployment shape. The L78 peak is empirically suspicious, however: it sits at the extreme end of the layer sweep in a region we do not expect to host

the cleanest deception representations, and the same metric on MASK val (Cell A) produces a quite different peak at L22. We read the L78 result as the signal exploiting an interaction between Alpaca and Tax val's in-distribution character rather than a general property of the probes.

MASK F1@1%FPR (Cell A). Best HP varies from 0.24 to 0.72 across layers (range 0.48), peaking at L32 with a broad peak structure across the high 20s and low 30s. Strong discrimination, deployment-shaped metric, naturalistic val set, peak in a layer region (early-mid) consistent with prior probe-effectiveness findings. The 0.48 range is broadly comparable to Cell C's 0.77 range despite the two val sets being completely different, indicating that the magnitude of variation is shaped substantially by Alpaca rather than by the val signal alone, but the *shape* does differ across the two cells, so the val set is influencing which layers stand out. Passes both criteria.

MASK AUROC (Cell B). Best HP scores 0.60–0.73 across layers (range 0.127), peaking at L28. Discrimination is modest in absolute terms but clearly above Tax AUROC's plateau and ~10^3 noise-σ above the noise floor. The signal does not match deployment shape, but is a single-source measure of val discrimination. Passes criterion 1 weakly, is less aligned with criterion 2 than the F1 cells.

Decision. Cell A (MASK F1@1%FPR) emerges as the strongest candidate. It discriminates strongly, mirrors the deployment metric, uses a naturalistic out-of-distribution val set, and peaks in a layer region consistent with prior probe-effectiveness findings. Cell B (MASK AUROC) is the next-best alternative. It has cleaner signal, but weaker discrimination and less aligned with deployment.

We acknowledge limits. First, Cell A's variation is jointly shaped by Alpaca and MASK, so we cannot fully decompose what the F1 signal is selecting for. Second, the four signals peak at different layers (L32, L28, L78, L30), and we cannot adjudicate between them with eval data; the choice rests on the criteria above rather than demonstrated superiority at predicting downstream eval performance. Third, optimising against any single val set risks overfitting to that set's idiosyncrasies. This is a concern that applies to all four candidates and that further independent validation distributions would help address.

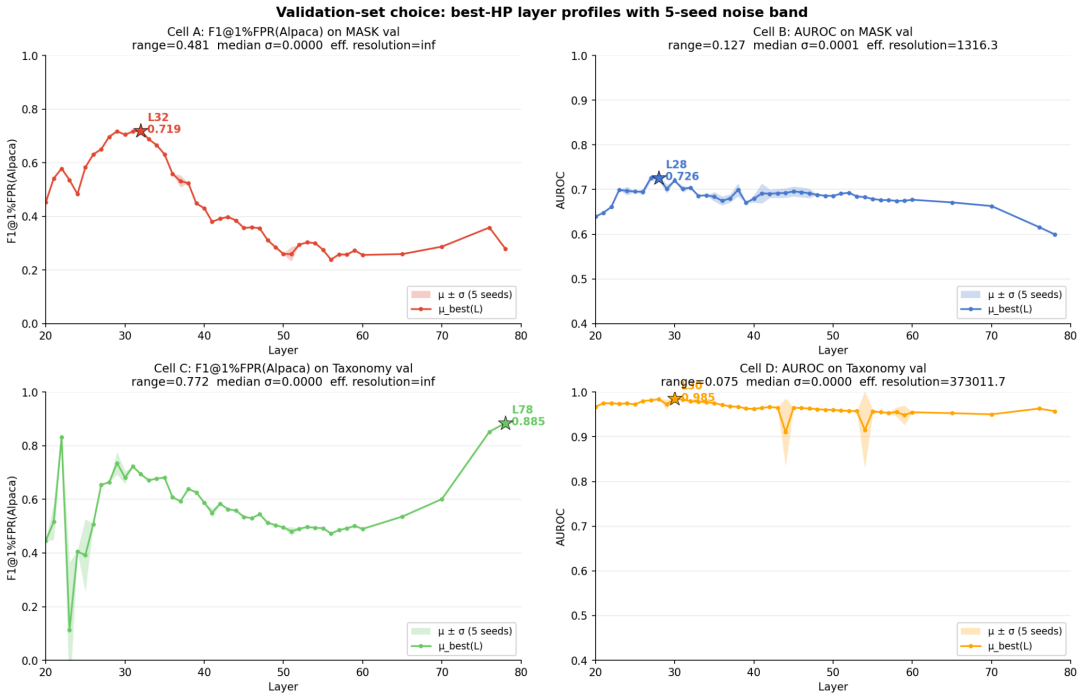

Figure 18

B: Best Apollo Probes Layer Sweep (All Metrics)

These plots show the full suite of metrics for the layer sweep of mean aggregated probes trained on the Apollo contrastive pairs dataset and selected by F1@1%FPR (Alpaca) on MASK.

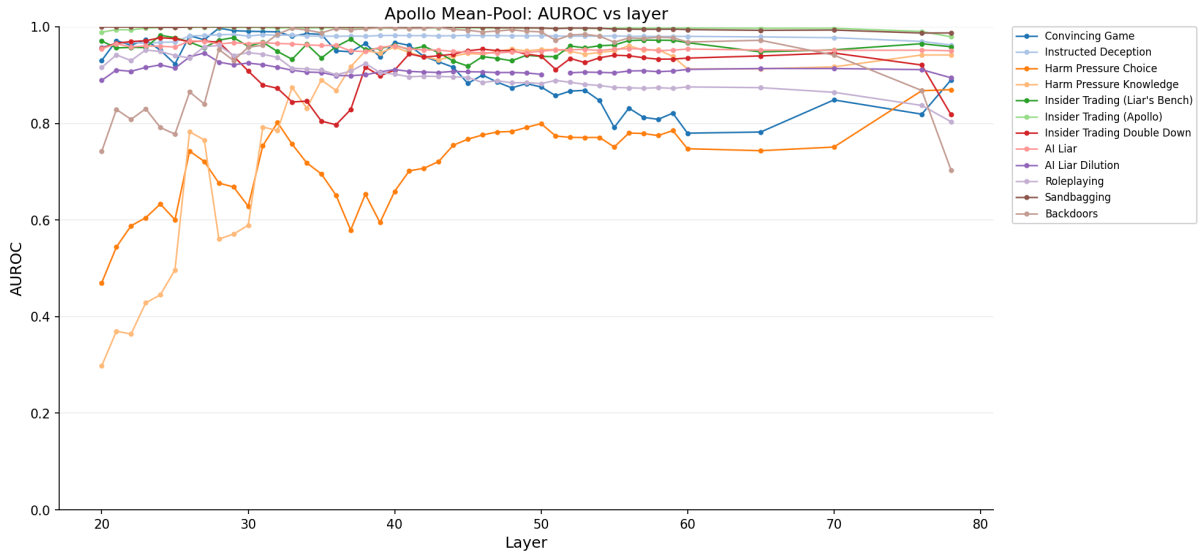

Figure 19

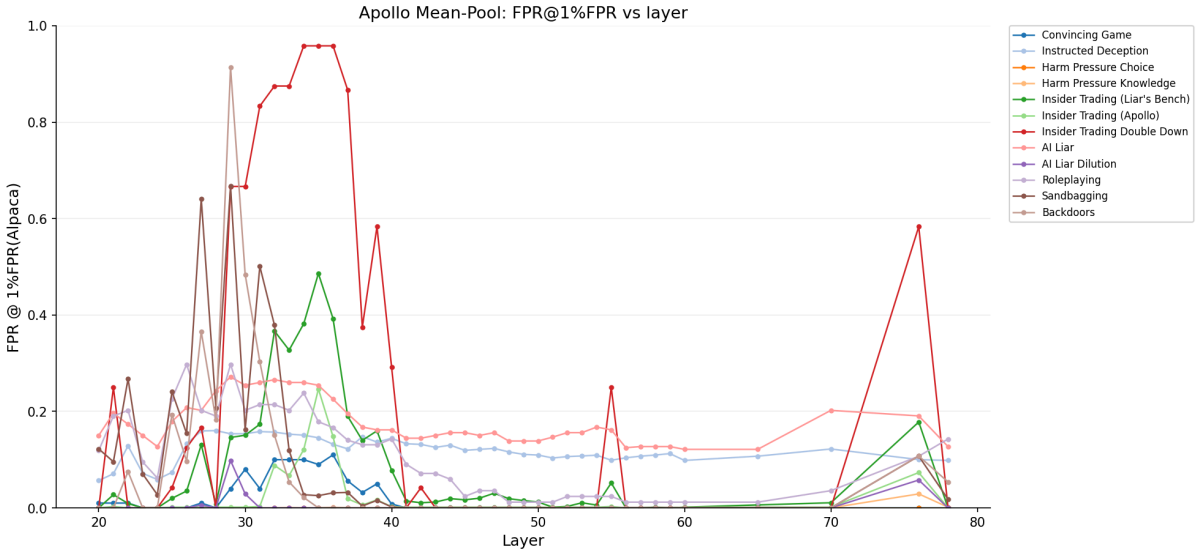

Figure 20

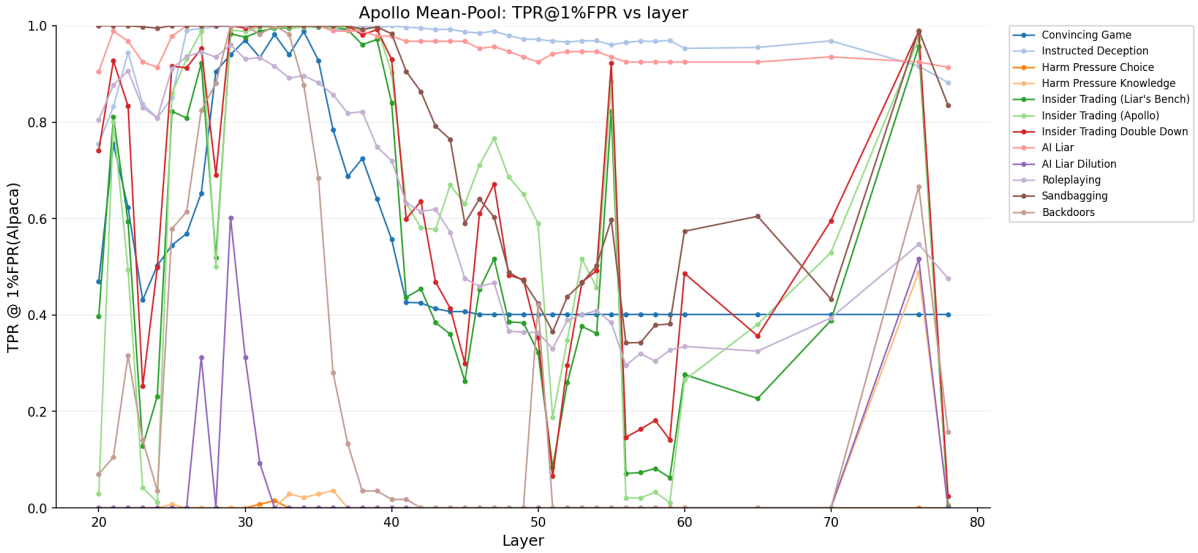

Figure 21

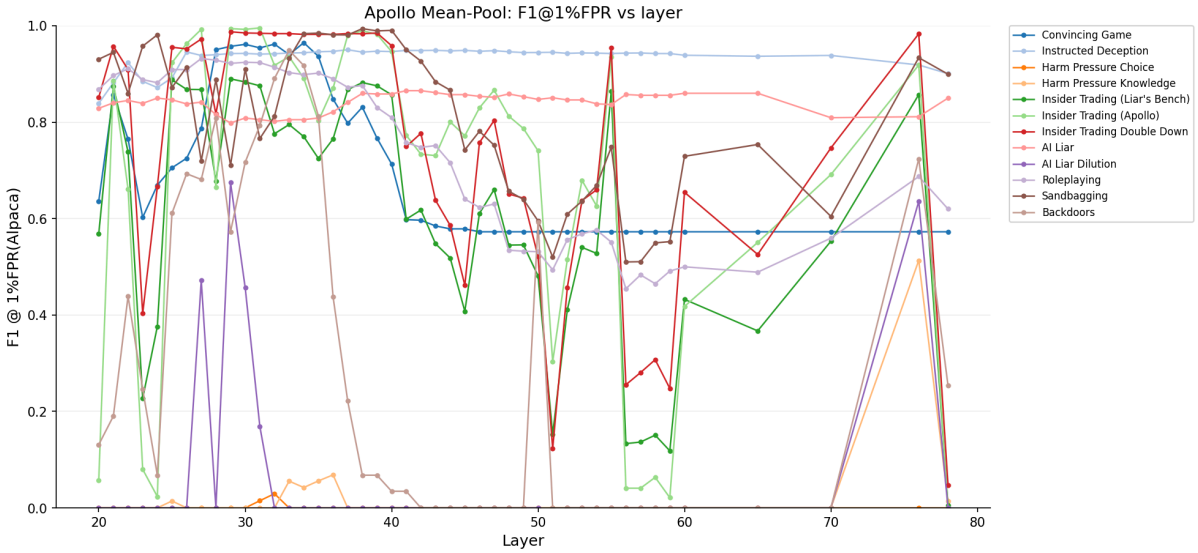

Figure 22

C: Probe vs BBR vs WBR (All Metrics)

These plots show the full suite of metrics for the Black Box and White Box Reasoning monitors using our best layer 32 probe and Gemma 4 E4B.

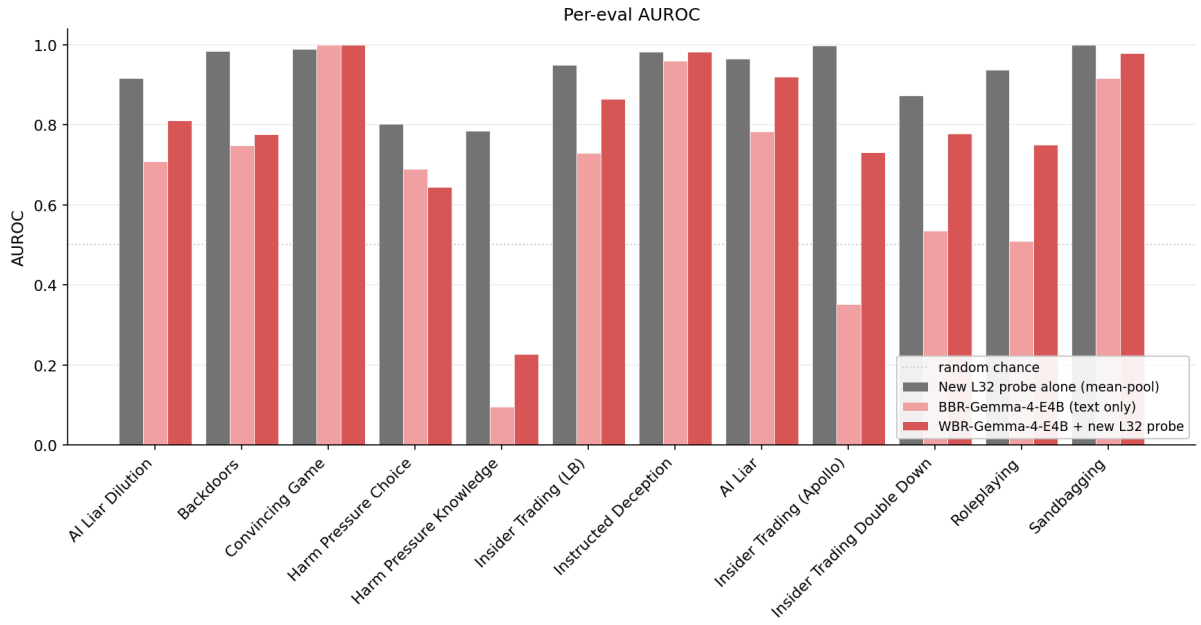

Figure 2

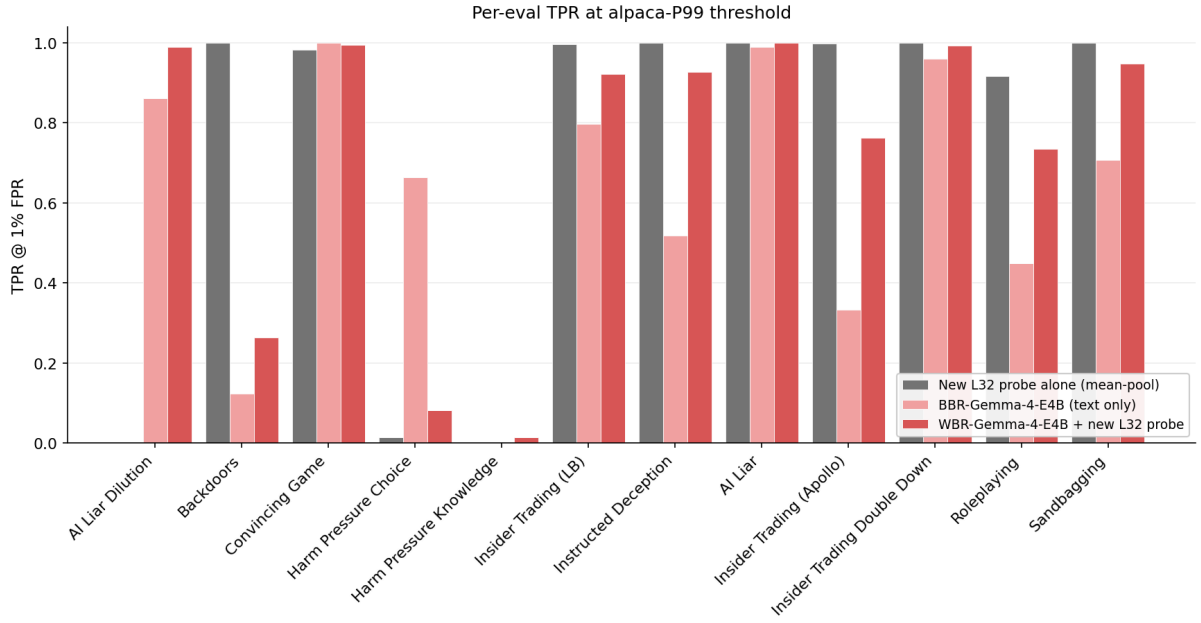

Figure 24

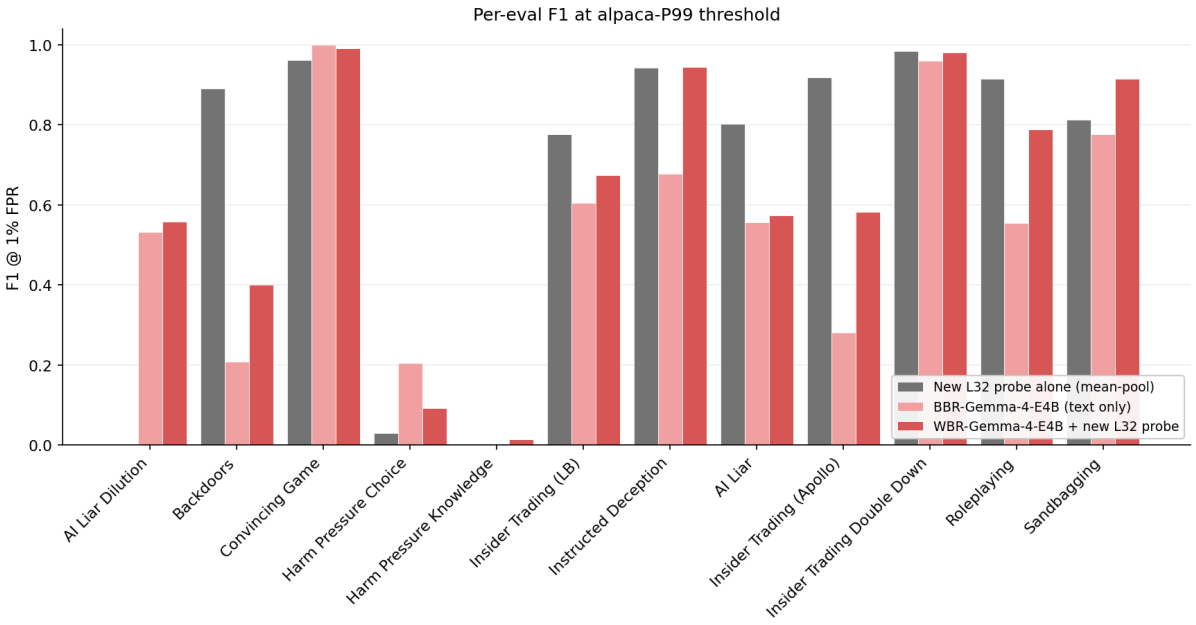

Figure 25

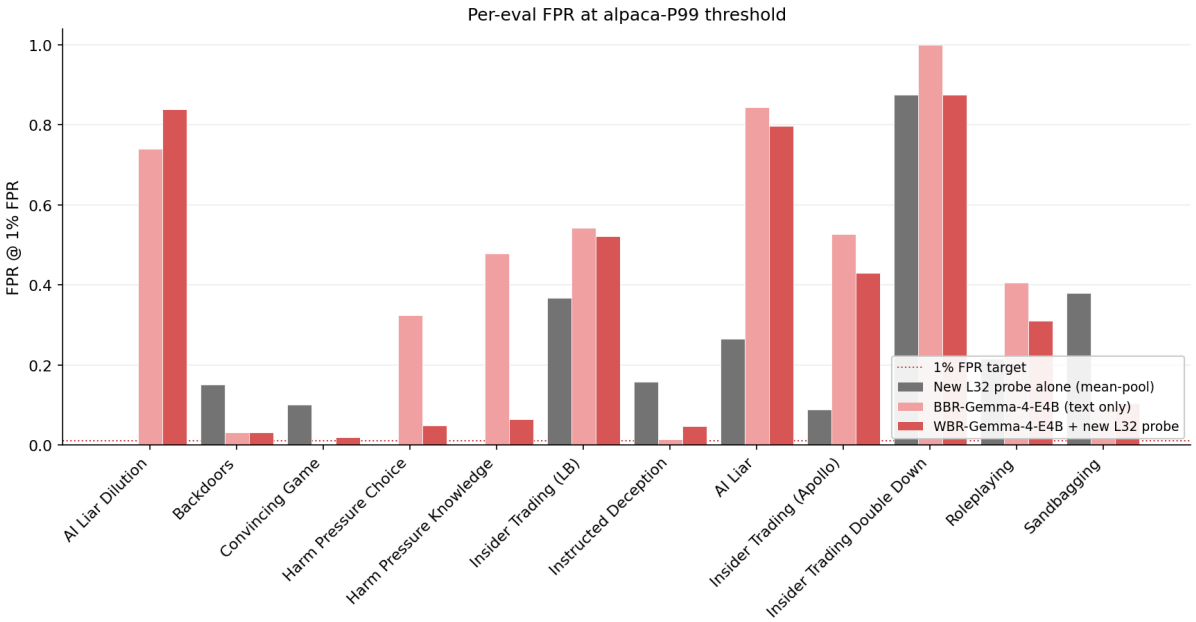

Figure 26

D: Additional Inter-Layer Metrics

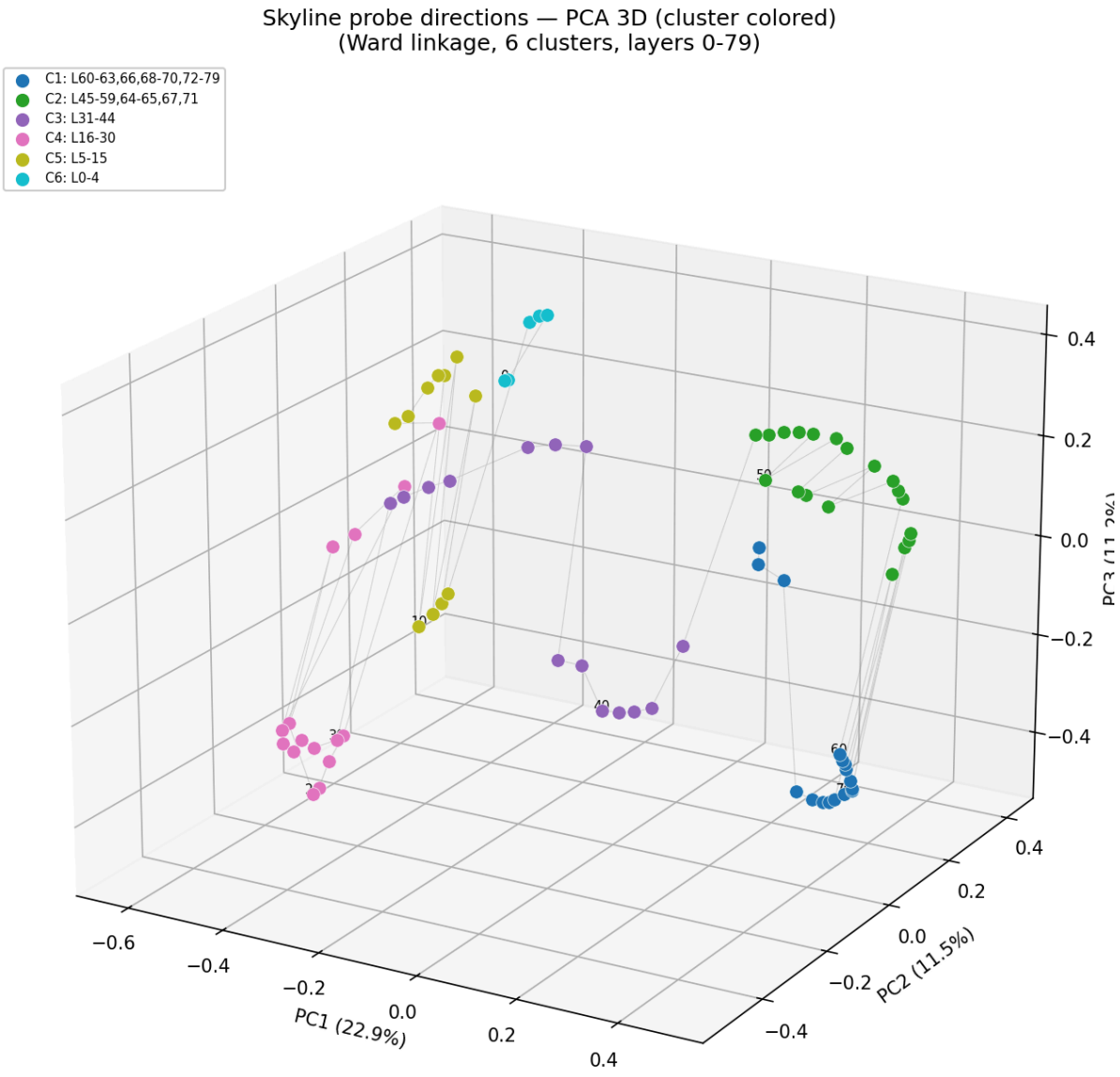

Figure 27

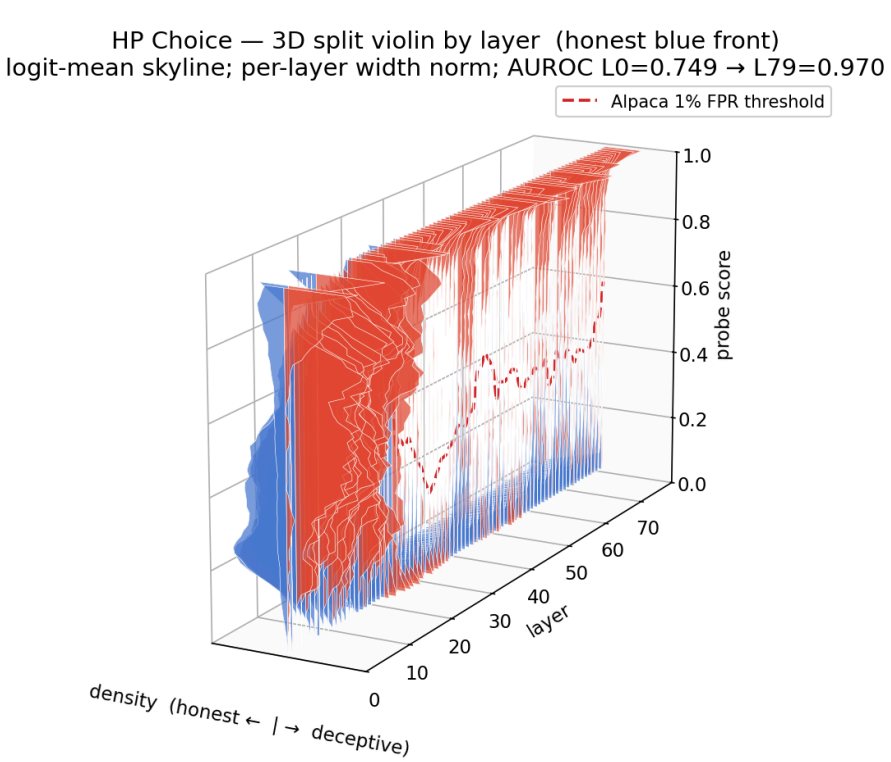

Figure 28

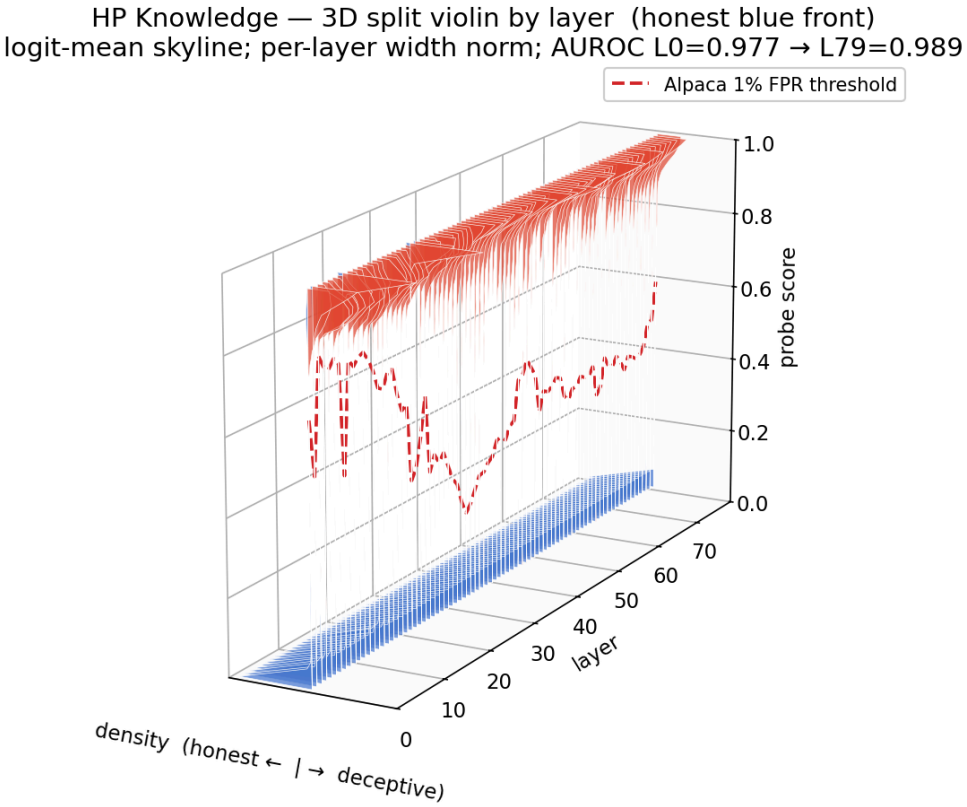

Figure 29

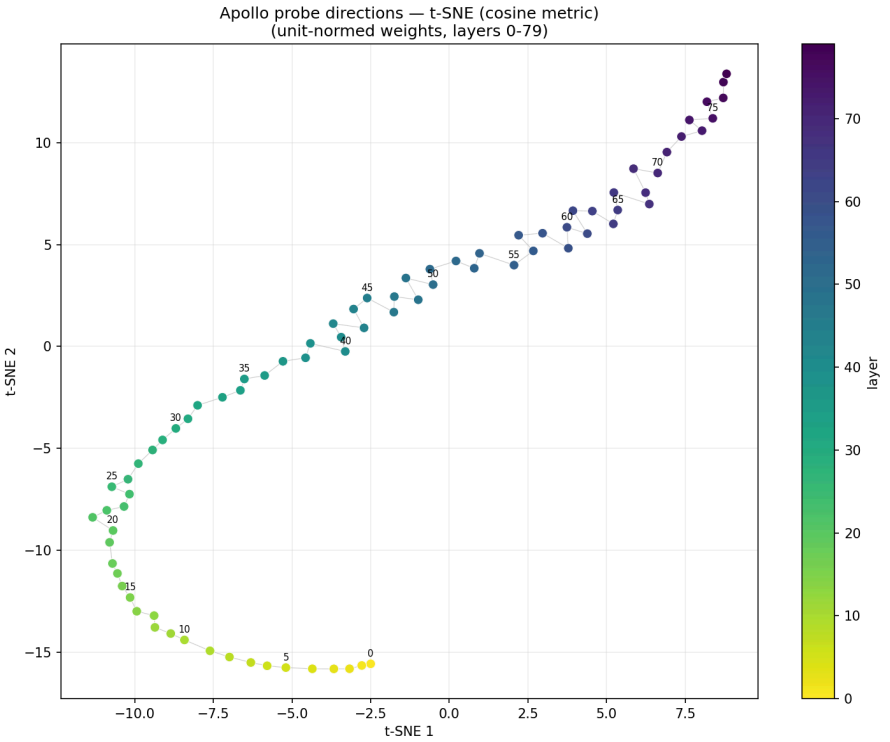

Figure 30

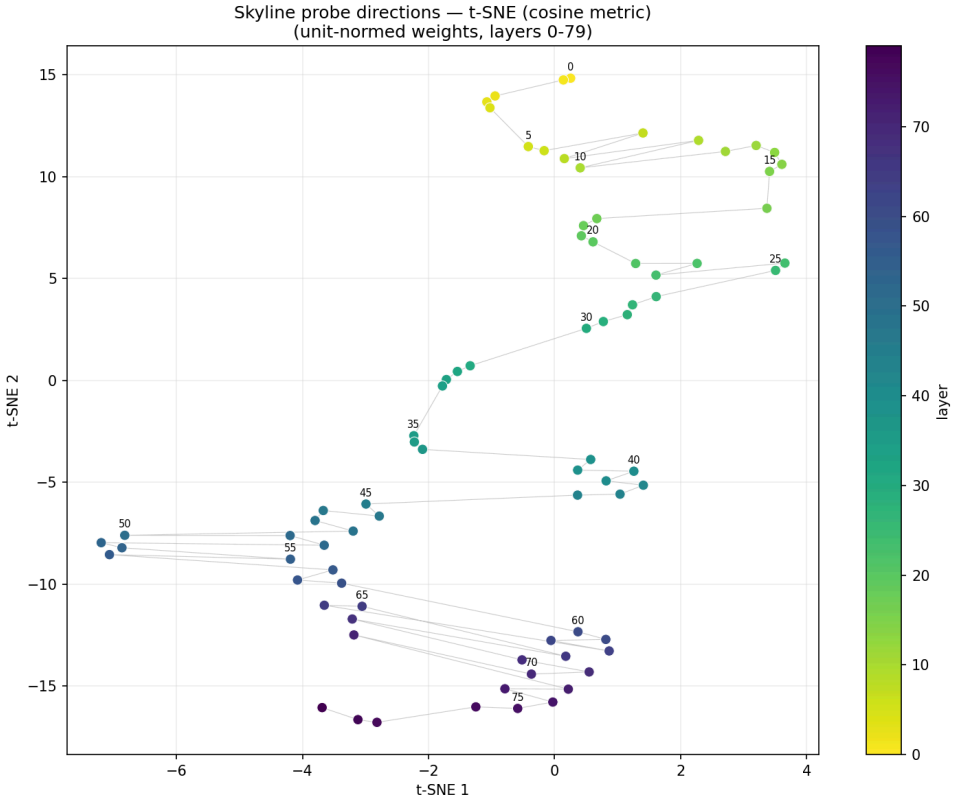

Figure 31