# Do RAG Hallucination Probes Need the Context? An Answer-Only Audit

**Atsushi Magata** [1] **Makoto Shimomura** [2]

## Abstract

Teacher-forced internal features—logit-lens trajectories and hidden-state dynamics—can strongly separate grounded answers from hallucinated hard negatives in retrieval-augmented generation benchmarks. However, high activation separability does not by itself imply that a detector uses retrieval context or implements a grounding-sensitive computation. We audit this interpretation by comparing answer-only (AO), query-answer (QWA), and query-context-answer (QCWA) teacher-forcing conditions, together with label-shuffle and hash-answer controls. On a stress-test set where Gemini-2.5-flash generates type-specific hard negatives conditioned on each existing (query, context, GT-answer) triple, AO nearly matches QCWA, exposing the failure mode in its most extreme form. On PHANTOM long-context financial QA, a hidden-state dynamics probe reaches AUPRC of 0.97 under QCWA, of which AO alone accounts for 80.6% of the above-baseline signal; the remaining QCWA−AO gap is small but non-zero, indicating a mixed regime rather than a pure artifact. On TruthfulQA, a short human-authored contrast set without retrieval context, AO separability is much weaker and QWA provides modest but consistent gains. Hash-answer and label-shuffle controls indicate that the AO signal is carried by the natural answer token sequence rather than by simple pipeline leakage. These results caution that hidden-state or logit-lens probes are not grounding detectors merely because they classify GT/HN answers under teacher forcing; AO audits are necessary to decompose answer-anchored and context-conditioned components.

[1]System And Data Engineering Group, Tokyo, Japan [2]Independent Researcher, Kanagawa, Japan. Correspondence to: Atsushi Magata <amagata3@sompo-japan.co.jp>, Makoto Shimomura <makoto.shimomura@kurikomi.jp>.

*Mechanistic Interpretability Workshop at the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

## 1. Introduction

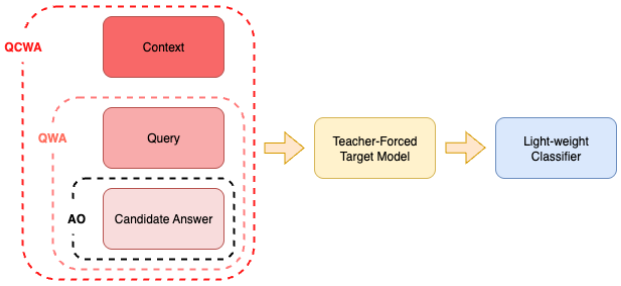


*Figure 1.* Answer-only audit protocol for teacher-forced internal detectors. The same candidate answer is fed under AO/QWA/QCWA to a teacher-forced target model; per-condition AUPRC comparisons decompose apparent detection into answer-anchored and context-conditioned components.

A hallucination probe that reads a model's internal activations at the answer-token positions can separate grounded from hallucinated answers *without ever using the retrieved context*: rather than checking whether the answer is supported by the context, it may be reading answer-side properties—length, likelihood, fluency, generator style, or dataset artifacts—off the answer tokens themselves. High activation separability under teacher forcing therefore does not, by itself, establish that a probe is a grounding detector. We make this failure mode measurable with a simple audit: feed the *same* candidate answer under three conditions—answer only (AO), query + answer (QWA), and query + context + answer (QCWA)—and compare per-condition separability. If AO already recovers most of the separability, the probe's success is largely answer-anchored rather than grounding-sensitive.

Teacher-forced (TF) hidden states, logit-lens features, and token-level dynamics appear to provide strong signals for distinguishing grounded from hallucinated answers (Orgad et al., 2025; Duan et al., 2024). This failure mode is grounded in prior findings that style and length can outweigh substance in model-based judgments (Feuer et al., 2025), and that labels can often be predicted from only one side of an input pair, as in hypothesis-only artifacts in NLI (Poliak et al., 2018; Gururangan et al., 2018; Geva et al.,

2019); recent work likewise suggests answer-anchored truthfulness signals in LLM internals (Luo et al., 2026). We ask whether an analogous failure mode affects mechanistic RAG hallucination detectors built on corpora such as PHANTOM (Ji et al., 2025) and natural RAG-hallucination resources (Niu et al., 2024).

**Contributions.** Our contribution is an *answer-only audit protocol* for teacher-forced RAG hallucination detectors, combining AO/QWA/QCWA comparisons with label-shuffle and hash-answer controls. We use it to show that GT-conditioned, Gemini-generated synthetic hard negatives can yield near-perfect internal-feature separability that is largely answer-only. We then apply the same audit to two public benchmarks: PHANTOM (Ji et al., 2025) exhibits a mixed regime with a large AO component plus a real QCWA context gain, whereas TruthfulQA (Lin et al., 2022) shows weaker AO shortcuts, consistent with short human-authored answers.

**Scope.** Our audit targets probes that read internal features *at or after* the answer-token positions. We distinguish three regimes: (i) *pre-generation* probes that see only the query and context and never the answer tokens; (ii) *during-generation* probes that see previously generated answer tokens; and (iii) *teacher-forced or post-hoc* probes that ingest the full candidate answer. The answer-only audit is most directly a critique of (iii), partially relevant to (ii), and *not* a critique of (i), where no answer tokens are available to anchor on. Throughout, "RAG hallucination probe" should be read in this restricted sense.

## 2. Audit Protocol and Diagnostic Predictions

**Diagnostic predictions.** We use three predictions: ($\mathrm{H}_{\mathrm{GROUNDING}}$) if TF features detect context grounding, QCWA should substantially outperform AO; ($\mathrm{H}_{\mathrm{ANSWER}}$) if they detect answer-side artifacts, AO should approach QCWA; and ($\mathrm{H}_{\mathrm{LEAKAGE}}$) if a high AO signal is due to pipeline leakage, label-shuffle or hash-answer controls should remain above chance.

**Input conditions and notation.** AO = Answer only; QWA = Query + Answer; QCWA = Query + Context + Answer. GT, HN, and *HS* denote ground-truth answers, hard-negative answers, and residual-stream hidden states (prefix in "HS geometry" / "HS dynamics"). Each slice has $n_l$ GT and $n_h$ HN examples; we report macro AUPRC with chance equal to the positive-class prevalence $\pi_{\text{pos}} = n_h/(n_h + n_l)$. All evaluations are 1 : 1 paired, so $\pi_{\text{pos}} = 0.5$. Decomposition metrics used in result tables: *gain* $(X{-}Y) = \mathrm{AUPRC}(X) - \mathrm{AUPRC}(Y)$ and *AO fraction* $= (\mathrm{AUPRC}(\mathrm{AO}) - \pi_{\text{pos}})/(\mathrm{AUPRC}(\mathrm{QCWA}) - \pi_{\text{pos}})$ measures the share of above-chance separability attainable from the answer alone: 1.0 means AOmatches QCWA's aggregate AUPRC, 0.0 means AOis at chance while QCWAis informative, and intermediate values indicate that QCWAexceeds AOin aggregate separability. Values above 1.0 (AO QCWA) are possible and indicate that adding context degrades aggregate separability. AUPRC values are macro averages across query-grouped CV folds (LogReg-L2 unless noted).

### 2.1. Feature families

Teacher forcing here is a *single-pass* evaluation: the prompt and candidate answer tokens are fed together as the observed sequence rather than generated autoregressively, and per-position hidden states and log-probabilities are read off at the answer-token positions in that one pass.

We extract TF features over the answer span and group them into trivial-confound baselines, an LM-head-free geometric family, and an incremental logit-lens / HS-dynamics ablation. The full feature inventory and per-family feature counts are deferred to Appendix A; we summarize here only what is needed to read the result tables.

For each layer $\ell$ and answer position $t$ we compute five per-layer metrics: three *logit-lens metrics* (logprob, entropy, gap; LM head used) and two *geometric metrics* ($\delta h$-norm, cos-sim; residual stream only, no LM head). Each metric is summarized along two axes: a *layer-axis aggregation* (final-layer scalar, full/all-layer centroid/peak/variance, last-valid layer, and early/mid/late group means) and, from Case 2 onwards, a *token-axis aggregation* for logit-lens metrics only (std / p10 / p90 / adjdiff over answer-token positions).

We use four trivial-confound baselines—**Ctrl A** = Length only (answer-token count), **Ctrl B** = log-PPL (final-layer log-probability, equal to $-\log$ Perplexity), **Ctrl C** = Length + log-PPL, **Ctrl D** = HS geometry only ($\approx$ 22 layer-axis aggregations of $\delta h$-norm and cos-sim, never projecting to vocabulary)— together with a four-step strictly cumulative ablation: **Case 0** = final-layer logit lens (3 features), **Case 1** = + multi-layer aggregations of logprob/entropy/gap (33 features), **Case 2** = + token-axis dynamics ($\approx$ 337 features), **Case 3** = Case 2 $\cup$ Ctrl D = "HS dynamics" ($\approx$ 365 features). Case 3 is the smallest configuration combining all three observation axes (layer $\times$ token $\times$ residual-stream geometry); it serves as the primary HS-dynamics setting in the AO/QWA/QCWA ablations below. AUPRC $> \pi_{\text{pos}}$ for Ctrl D therefore implies signal independent of the language head.

## 3. Datasets and Target Models

**A. GT-conditioned Gemini HNs.** We construct a stress-test set by prompting Gemini-2.5-flash to generate four type-specific hard-negative answers for each existing (query, context, GT-answer) triple sourced from PHANTOM (con-

tent / qualification / structural / retrieval-grounding error types). The queries, contexts, and GT answers are inherited from the source RAG dataset; only the HN answers are newly generated. We refer to these as GT-*conditioned Gemini* HN*s* (full construction protocol in Appendix F). This setting exposes an extreme answer-only shortcut.

**B. PHANTOM official (Ji et al., 2025).** PHANTOM generates faithful and hallucinated financial answers from 500-token chunks of U.S. Securities and Exchange Commission (SEC) filings using Llama-3.3-70B-instruct, followed by expert validation, and later expands context length and placement variants. For our audit we use the long-context variants, selecting examples whose GPT-OSS-20B-tokenized context lengths fall in the 7K–14K token range (median $\approx$ 11.2K, p10–p90 $\approx$ 10.3K–12.9K; see Appendix D), rather than the original 500-token seed chunks. This places the audit in a genuinely long-context RAG regime. Paired GT and HN samples derived from the same seed example are grouped together in CV folds so that the same seed never appears in both train and test. PHANTOM reduces generator-identity confounding compared with our Gemini-HN data, but still may contain faithful-vs-hallucinated generation-mode artifacts (Geva et al., 2019). Compared with naturally generated RAG-hallucination corpora such as RAGTruth (Niu et al., 2024), PHANTOM's synthetic faithful/hallucinated dichotomy makes the answer-side artifact risk we audit more salient.

**C. TruthfulQA (Lin et al., 2022).** TruthfulQA provides a human-authored, short-answer contrast set. Because it lacks full retrieval context, we use it only for AO/QWA truthfulness separability, not RAG grounding.

**Target Models for Teacher forcing.** GPT-OSS-20B and Qwen3-4B-AWQ (AWQ = activation-aware weight quantization).

## 4. Main results

**Result 1: GT-conditioned Gemini HNs are almost entirely answer-only.** In the most artifact-prone setting, AO nearly matches QCWA at every stage of the ablation. Hashing answers collapses every AO/QWA feature to $\approx \pi_{\text{pos}}$ and within-query label-shuffle collapses all Cases to chance (Appendix B); the AO signal is therefore carried by the answer token sequence rather than by pipeline leakage.

**Result 2: PHANTOM is a mixed regime, not a pure artifact.** QCWA dominates QWA and AO across all ablation stages, but AO already recovers $80.6\%$ of the above-baseline QCWA separability at Case 3 (AO fraction $= 0.806$; Table 1). Length-only is non-trivial ($\approx 0.73$), reflecting the GT/HN length asymmetry of the PHANTOM generator. The same qualitative pattern holds under Qwen3-4B-AWQ (Appendix C).

**Result 3: TruthfulQA shows weaker answer-only shortcuts.** Length-only collapses to $\approx 0.53$, eliminating the dominant length confound seen on PHANTOM. QWA exceeds AO consistently from Case 1 onward ($\approx +0.02$ to $+0.08$; Table 2), with the largest gains in the richer Case 2 / 3 settings, especially for Qwen3-4B-AWQ. The pattern is consistent across GPT-OSS-20B and Qwen3-4B-AWQ (Appendix C).

**Take-aways for PHANTOM.** QCWA strictly dominates AO at Case 3 ($+0.092$), but AO recovers $80.6\%$ of the above-baseline QCWA separability (AO fraction $= 0.806$)—most of the discriminability over $\pi_{\text{pos}}$ is already attainable without ever seeing the question or context. The gain (QCWA$-$QWA) is comparable in magnitude to gain (QCWA$-$AO), indicating a meaningful context-conditional component on top of the AO shortcut.

**Why the AO control matters.** Without it, PHANTOM Case 3 would appear to provide near-perfect evidence of grounding-sensitive detection (QCWA AUPRC$= 0.970$). With AO we find that AUPRC$= 0.879$ is already attainable without query or context, accounting for $80.6\%$ of the above-baseline separability. The remaining QCWA$-$AO gain shows that PHANTOM is *not* a pure answer-only artifact, but the high QCWA score alone is not identifiable as grounding: the same QCWA AUPRC is compatible with two incompatible interpretations (context grounding vs. answer-side separability) unless AO is reported alongside it.

**Take-aways for TruthfulQA.** Length-only is near chance ($\approx 0.53$), much weaker than on PHANTOM. AO still reaches $\approx 0.64$ at Case 3 (both target models), indicating residual answer-side artifacts in human-authored distractors, but QWA exceeds AO at every Case from Case 1 onwards ($\approx +0.03 / +0.08$). The pattern is consistent across both target models, indicating that on TruthfulQA a non-trivial portion of the QWA signal reflects query-conditioned answer compatibility rather than answer-only cues or RAG context grounding, which is unavailable here. TruthfulQA changes two factors at once relative to PHANTOM—answers are *both* human-authored and much shorter—so the weaker AO separability should not be attributed to human authorship alone; it is also consistent with the hypothesis that long synthetic explanatory answers amplify answer-side separability.

## 5. Discussion

**Activation separability is not grounding.** A hidden-state probe that separates GT and HN under QCWA is *not*, on its own, evidence that the probe is a grounding detector.

*Table 1.* PHANTOM QCWA/QWA/AO decomposition. GPT-OSS-20B target model, LogReg-L2, $n_q = 396$ paired GT vs. PHANTOM-original HN. Column definitions (*gain*, *AO fraction*): see notation in Section 2.

| Feature | QCWA | QWA | AO | gain (QCWA−AO) | gain (QCWA−QWA) | AO frac. |
|---|---|---|---|---|---|---|
| Ctrl A: Length only | 0.729 | 0.729 | 0.729 | 0.000 | 0.000 | 1.000 |
| Ctrl B: log-PPL | 0.815 | 0.473 | 0.589 | +0.226 | +0.342 | 0.282 |
| Ctrl C: Length + log-PPL | 0.835 | 0.729 | 0.727 | +0.108 | +0.107 | 0.678 |
| Ctrl D: HS geometry only | 0.931 | 0.813 | 0.852 | +0.079 | +0.118 | 0.817 |
| Case 0: Final logit lens | 0.881 | 0.583 | 0.638 | +0.243 | +0.298 | 0.362 |
| Case 1: + Multi-layer | 0.913 | 0.781 | 0.768 | +0.145 | +0.132 | 0.649 |
| Case 2: + Token-axis | 0.969 | 0.898 | 0.867 | +0.101 | +0.071 | 0.783 |
| Case 3: HS dynamics | 0.970 | 0.902 | 0.879 | +0.092 | +0.068 | 0.806 |

*Table 2.* TruthfulQA QWA/AO decomposition. LogReg-L2, $n_q = 817$ GT/incorrect pairs. QCWA and *AO fraction* omitted: TruthfulQA has no retrieval context, so no QCWA ceiling. gain = AUPRC(QWA)−AUPRC(AO).

| | GPT-OSS-20B | | | Qwen3-4B-AWQ | | |
|---|---|---|---|---|---|---|
| Feature | QWA | AO | gain | QWA | AO | gain |
| Ctrl A: Length only | 0.534 | 0.534 | 0.000 | 0.534 | 0.534 | 0.000 |
| Ctrl B: log-PPL | 0.477 | 0.517 | −0.039 | 0.527 | 0.508 | +0.019 |
| Ctrl D: HS geometry only | 0.647 | 0.643 | +0.005 | 0.614 | 0.613 | +0.001 |
| Case 0: Final logit lens | 0.552 | 0.513 | +0.039 | 0.517 | 0.526 | −0.008 |
| Case 1: + Multi-layer | 0.619 | 0.588 | +0.031 | 0.663 | 0.593 | +0.070 |
| Case 2: + Token-axis | 0.650 | 0.618 | +0.032 | 0.711 | 0.635 | +0.076 |
| Case 3: HS dynamics | 0.664 | 0.641 | +0.023 | 0.715 | 0.638 | +0.077 |

If the same probe achieves comparable separability under AO—where the prompt and the retrieved context have been removed and only the answer span is available—then its discriminative behavior is consistent with exploiting *answer-anchored structure in the residual stream* rather than a *context–answer support relation*. The QCWA AUPRC, taken alone, cannot distinguish these two regimes. High QCWA probe performance is compatible with at least four mechanisms: genuine context-grounding computation, answer-only textual artifacts (Poliak et al., 2018; Gururangan et al., 2018), generation-mode artifacts such as faithful vs. hallucinated prompt mode under a single generator (Geva et al., 2019), and likelihood / length / style shortcuts (Feuer et al., 2025). Only the first is grounding-sensitive; the latter three are answer-anchored and survive removal of query and context, so they show up under AO. The AO/QWA/QCWA decomposition reported in Tables 1–2 separates the contribution of grounding (the QCWA−AO gain) from the answer-anchored mechanisms (the AO baseline above $\pi_{\text{pos}}$).

**AO control is a decomposition tool, not just a rejection test.** On self-generated HN, separability is dominated by AO. On PHANTOM, AOis supplemented by a measurable context gain. On TruthfulQA, the AObaseline is weaker and a modest query gain emerges.

**Practical implications.** The audit also suggests concrete mitigations for practitioners building or evaluating teacher-forced RAG hallucination detectors. First, AO (and, where retrieval context exists, QCWA−AO) should be reported alongside any headline QCWA number, so that the grounding-sensitive component is identifiable rather than conflated with answer-anchored separability. Second, when constructing hard-negative benchmarks, controlling answer length and style between GT and HN (as in our $\pm 30\%$ matching constraint) limits the dominant length confound. Third, where the deployment goal is to assess grounding before commitment to an answer, pre-generation probes that never see the answer tokens sidestep the answer-anchored shortcut entirely. We frame these as evaluation discipline rather than a new detector: the contribution is the audit, not a replacement probe.

**Dataset construction matters.** PHANTOM (Ji et al., 2025): long, synthetic faithful/hallucinated answers; strong length cue. TruthfulQA (Lin et al., 2022): short, human-authored answers; length cue weak. For naturally generated RAG hallucinations (e.g. RAGTruth (Niu et al., 2024)) the nature of answer-side distribution shift may differ from synthetic faithful-vs-hallucinated construction; our audit protocol still applies and is necessary.

**Relation to Two Pathways and internal hallucination signals.** Recent work argues for answer-anchored and question-anchored truthfulness pathways (Luo et al., 2026), and shows that LLM hidden states encode hallucination/truthfulness signals that simple probes can read out (Orgad et al., 2025; Duan et al., 2024). Our contribution is not the discovery of answer-anchored signals. Rather, we show that such signals create a concrete evaluation failure

mode for mechanistic RAG hallucination detectors: high activation separability can be largely answer-only even under query-grouped CV.

**A note on attention as evidence.** We include qualitative QCWA-side attention case studies in Appendix E but treat them as illustrative only: the aggregate interpretation comes from AO/QWA/QCWA controls rather than from attention peaks, since per-instance attention patterns turn out to be heterogeneous (sometimes localized on seed-context, sometimes diffuse) even within the same probe regime.

## 6. Limitations

Our study has several limitations. TruthfulQA lacks full retrieval context and therefore serves only as a short-answer truthfulness contrast, not as a RAG grounding benchmark. PHANTOM uses synthetic generation modes, so its AO separability may partly reflect faithful-vs-hallucinated prompt artifacts. We do not claim that context signal is absent—PHANTOM shows a clear QCWA−AO gain—nor that internal features are uninformative; rather, our claim is interpretive: such features require AO/QWA/QCWA controls before being described as grounding-sensitive. Finally, human-vs-synthetic and short-vs-long answer effects are not fully disentangled in our two public benchmarks.

## Code and Data Availability

Code for feature extraction, probe training, and the AO/QWA/QCWA decomposition—together with the Gemini hard-negative generation prompts, the dataset splits, and the random seeds used in all experiments—is available at https://github.com/amagata-dx/tf-probe.

## Impact Statement

This paper presents work whose goal is to advance the methodology of evaluating mechanistic LLM hallucination detectors in retrieval-augmented generation. Stronger evaluation discipline can reduce the risk that systems are deployed on the basis of spurious internal-feature separability rather than genuine context-grounding. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## References

Duan, H., Yang, Y., and Tam, K. Y. Do LLMs know about hallucination? an empirical investigation of LLM's hidden states, 2024.

Feuer, B., Goldblum, M., Datta, T., Nambiar, S., Besaleli, R., Dooley, S., Cembalest, M., and Dickerson, J. P. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)*, 2025.

Geva, M., Goldberg, Y., and Berant, J. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, 2019.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, 2018.

Ji, L., Seyler, D., Kaur, G., Hegde, M., Dasgupta, K., and Xiang, B. PHANTOM: A benchmark for hallucination detection in financial long-context QA. In *Advances in Neural Information Processing Systems*, 2025. Datasets and Benchmarks Track.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, 2022.

Luo, W., Peng, G., Li, W., Wei, S., Song, F., Wang, L., Yang, N., Zhang, X., Jin, J., Wei, F., and Wang, H. Two pathways to truthfulness: On the intrinsic encoding of LLM hallucinations, 2026. Accepted to ACL 2026 Main Conference.

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., and Zhang, T. 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 2024)*, 2024.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)*, 2025.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018)*, 2018.

# A. Feature Inventory and Implementation Details

## A.1. Feature inventory

Combined inventory of the four trivial-confound baselines (Ctrl A–D) and the four-step cumulative ablation (Cases 0–3). Feature counts are for GPT-OSS-20B (25 layers).

*Table 3.* Feature inventory for the trivial-confound baselines (Ctrl A–D) and the cumulative ablation (Cases 0–3) used in the AO/QWA/QCWA decomposition. Counts are for GPT-OSS-20B (25 layers).

| Family | Components | $n_{\text{feat}}$ | Notes |
|---|---|---|---|
| Ctrl A: Length only | answer-token count ($n_a$) | 1 | Answer token count under the TF tokenizer. |
| Ctrl B: log-PPL | final-layer mean log-prob | 1 | Mean log-prob of the answer span at the final layer ($= -\log$ Perplexity; log-scale, outlier-robust). |
| Ctrl C: Length + log-PPL | both above | 2 | Joint trivial-confound baseline. |
| Ctrl D: HS geometry only | layer-axis aggregations of $\delta h$-norm and cos-sim | ≈22 | Residual-stream geometric only; never projects to vocabulary. |
| Case 0: Final logit lens | final-layer logprob / entropy / gap | 3 | Final-layer scalars only. |
| Case 1: + Multi-layer | all layer-axis aggregations of logprob/entropy/gap | 33 | Adds full / all / last-valid / early / mid / late aggregations on top of Case 0. |
| Case 2: + Token-axis | token-axis stats (std / p10 / p90 / adjdiff) over Case 1 features | ≈337 | Adds answer-token-position distribution statistics. |
| Case 3: HS dynamics | $\delta h$-norm and cos-sim layer aggregations on top of Case 2 | ≈365 | Case 2 ∪ Ctrl D ("HS dynamics minimal"); primary HS-dynamics setting in the AO/QWA/QCWA ablations. |

## A.2. Case definitions and feature counts

Each Case strictly extends the previous; with set-theoretic abuse of notation we identify a Case with its underlying feature set:

- Case 0 = {final-layer logprob, entropy, gap}, $|\text{Case 0}| = 3$.
- Case 1 = Case 0 ∪ {layer-axis aggregations of logprob / entropy / gap}, $|\text{Case 1}| = 33$.
- Case 2 = Case 1 ∪ {token-axis stats (std, p10, p90, $\text{adjdiff}_{\text{max}}$, $\text{adjdiff}_{\text{mean}}$) over Case 1's per-layer logit-lens features}, $|\text{Case 2}| \approx 337$.
- Case 3 = Case 2 ∪ Ctrl D = Case 2 ∪ {layer-axis aggregations of $\delta h$-norm and cos-sim}, $|\text{Case 3}| \approx 365$.

Feature counts assume GPT-OSS-20B (25 layers); they vary with the per-model count of full-attention layer subsets and with whether the "last-valid-layer" aggregation is well-defined for each metric (e.g. $\delta h$-norm is undefined at the final layer).

## A.3. Notes on log-PPL, raw perplexity, and outliers

We use the final-layer mean log-prob (log-scale, equal to $-\log$ Perplexity) rather than raw perplexity $\exp(-\text{logprob})$ because raw perplexity has heavy-tailed outliers that break StandardScaler-based linear probes. On Qwen3-4B-AWQ × PHANTOM GT ($n_q = 396$) the AWQ-quantized LLM produces sporadic near-zero token probabilities, yielding raw perplexity mean $\approx 1.1 \times 10^{15}$ and std $\approx 2.3 \times 10^{16}$, while the median remains $5.89$. Mean and std are dominated by a handful of tokens whose log-prob is near machine-zero; log-PPL is the standard log-scale fix and is the one we use throughout the paper. Under any monotone-invariant probe (e.g. XGBoost) raw perplexity and log-PPL are equivalent to Ctrl B; we use the log-scale form to keep linear probes well-conditioned across all target models.

### A.4. Probe configurations

All probes are scikit-learn / XGBoost implementations trained on the feature vectors defined in Appendix A. Numerical features are standardized with scikit-learn's StandardScaler before the linear and MLP probes (XGBoost is invariant to monotone feature transforms). Hyperparameters are fixed across all experiments rather than tuned per-Case, to avoid inflating the Case-3 AUPRC by hyperparameter selection on the same evaluation slice.

**LogReg-L2.** L2-regularised logistic regression (scikit-learn's LogisticRegression, $\ell_2$ penalty) with $C = 1.0$. This is the primary probe used in main Tables 1–2; we choose a linear probe so that "the probe achieves separability" is a statement about a low-capacity readout from the residual stream rather than about the capacity of the probe itself.

**XGBoost.** Gradient-boosted decision trees (XGBoost's XGBClassifier) with max-depth $= 3$, n-estimators $= 50$, and learning-rate $= 0.05$. Used in Appendix C as a flexible non-linear probe to check that the qualitative AO/QWA/QCWA pattern is not specific to linear readout.

**MLP.** Feed-forward neural network (scikit-learn's MLPClassifier) with two hidden layers of sizes $(128, 64)$, ReLU activations, and L2 weight decay $\alpha = 10^{-3}$. Used in Appendix C as a third cross-check.

**Cross-validation.** All AUPRC numbers are macro averages over a query-grouped 5-fold outer CV with 3-fold inner CV for any per-fold normalisation; the same query never appears in both train and test of the outer split.

## B. Sanity Checks (GT-conditioned Gemini HNs)

We verify that the high original AUPRC is not due to direct label leakage or non-textual pipeline artifacts. Two controls are applied to the same eval frame: (a) **Shuffle**—within-query permutation of the binary label $y$ before CV; (b) **Hash**—replacing each answer text with its SHA-256 prefix (16 chars) prior to teacher forcing. Both controls should bring AUPRC to chance ($= \pi_{\text{pos}}$). All numbers in this section are macro AUPRC under LogReg-L2.

*Table 4.* Sanity-check macro AUPRC on the GT-conditioned Gemini-HN benchmark. Original is the unmodified eval; Shuffle permutes the binary label $y$ within `query_id` before CV; Hash replaces each answer with a fixed-format SHA-256 prefix prior to teacher forcing. Both controls collapse every feature to $\approx \pi_{\text{pos}}$, indicating that the original AO signal is carried by the natural answer token sequence rather than direct label or pipeline artifacts.

| Feature | Cond. | Original | Shuffle | Hash |
|---|---|---|---|---|
| Ctrl A: Length only | – | 0.688 | 0.376 | 0.491 |
| Ctrl B: log-PPL | AO | 0.737 | 0.348 | 0.504 |
| Ctrl B: log-PPL | QWA | 0.543 | 0.404 | 0.508 |
| Ctrl C: Length + log-PPL | AO | 0.774 | 0.362 | 0.491 |
| Ctrl C: Length + log-PPL | QWA | 0.700 | 0.379 | 0.507 |
| Ctrl D: HS geometry only | AO | 0.971 | 0.403 | 0.478 |
| Ctrl D: HS geometry only | QWA | 0.971 | 0.400 | 0.512 |
| Case 0: Final logit lens | AO | 0.843 | 0.354 | 0.495 |
| Case 0: Final logit lens | QWA | 0.646 | 0.409 | 0.513 |
| Case 1: + Multi-layer | AO | 0.971 | 0.384 | 0.512 |
| Case 1: + Multi-layer | QWA | 0.960 | 0.392 | 0.488 |
| Case 2: + Token-axis | AO | 0.932 | 0.405 | 0.518 |
| Case 2: + Token-axis | QWA | 0.933 | 0.380 | 0.495 |
| Case 3: HS dynamics | AO | 0.944 | 0.400 | 0.518 |
| Case 3: HS dynamics | QWA | 0.934 | 0.377 | 0.496 |

### B.1. Label-shuffle control

The Shuffle column reports macro AUPRC after a single within-query permutation of the binary label $y$ before CV. Every Case (and every Ctrl) collapses to a no-skill regime around $\pi_{\text{pos}}$, with no learner exceeding $\approx 0.41$. The Case-by-Case AUPRC values do not exhibit a Case-progression pattern (i.e. additional features do not recover discriminative signal under shuffled labels), confirming that the Case-by-Case AUPRC progression in main Tables 1–2 is not exploitable under permuted labels.

### B.2. Hash-answer control

The Hash column reports macro AUPRC after replacing each answer text with a fixed-format SHA-256 hex prefix (16 characters) prior to teacher forcing. The character-level length is constant by construction; the tokenizer-level length depends on the target model's BPE vocabulary on hex strings and is approximately constant in practice (verifying exact tokenized-length invariance is model-specific). This control therefore substantially removes natural answer length/content variation while preserving pipeline structure. Every internal-feature AUPRC collapses to $\approx \pi_{\text{pos}}$, providing evidence against direct label/metadata leakage and showing that the original AO signal in main Tables 1–2 depends on the natural answer token sequence rather than positional or pipeline artifacts.

### B.3. Double-shuffle check

Single-label-shuffle values above dip below $\pi_{\text{pos}}$ for some Cases (e.g. Case 2/3 at AO $\approx 0.40$). This is not evidence of leakage: when probes are trained on randomly permuted labels, they may converge to directions anti-correlated with the original labels, and AUPRC measured against the still-original test labels reflects this anti-correlation. To disentangle this from leakage we ran an auxiliary check in which both the training and the evaluation labels are independently permuted within each query group. Under this double-shuffle, AUPRC converges to $\pi_{\text{pos}}$ within Monte-Carlo noise (50-seed mean $\approx \pi_{\text{pos}}$), confirming that the below-chance dip in the single-shuffle case is a within-group permutation artifact rather than a learnable signal.

## C. Cross-Target Model and Cross-Learner Robustness

### C.1. PHANTOM: Case-3 robustness across target models and learners

Headline AUPRC at Case 3 (HS dynamics = Case 2 $\cup$ Ctrl D, $\approx 365$ features; primary HS-dynamics setting). LogReg-L2 $\times$ GPT-OSS-20B is reproduced from main Table 1 for direct comparison.

*Table 5.* PHANTOM Case-3 (HS-dynamics) cross-target model and cross-learner robustness. Macro AUPRC for QCWA/ QWA/ AO and the corresponding QCWA-gains. The mixed-regime pattern (QCWA $>$ QWA $\geq$ AO) holds across both target models and the log-linear / MLP probes; the XGBoost $\times$ GPT-OSS-20B inversion is discussed below.

| Target model | Learner | QCWA | QWA | AO | gain (QCWA−AO) | gain (QCWA−QWA) |
|---|---|---|---|---|---|---|
| GPT-OSS-20B | LogReg-L2 | 0.970 | 0.902 | 0.879 | +0.092 | +0.068 |
| GPT-OSS-20B | XGBoost | 0.969 | 0.907 | 0.982 | −0.013 | +0.062 |
| GPT-OSS-20B | MLP | 0.963 | 0.819 | 0.869 | +0.094 | +0.144 |
| Qwen3-4B-AWQ | LogReg-L2 | 0.972 | 0.928 | 0.905 | +0.068 | +0.044 |
| Qwen3-4B-AWQ | XGBoost | 0.981 | 0.918 | 0.904 | +0.076 | +0.063 |
| Qwen3-4B-AWQ | MLP | 0.963 | 0.902 | 0.870 | +0.092 | +0.061 |

The XGBoost $\times$ GPT-OSS-20B row shows AO $>$ QCWA, suggesting that flexible nonlinear learners can exploit answer-only structure particularly strongly on this small-sample ($n_q = 396$) regime. We avoid over-interpreting this single inversion and use LogReg-L2 as the primary probe for the AO/QWA/QCWA decomposition. The log-linear and MLP probes agree that QCWA dominates AO by $\sim 0.07$–$0.09$, and both target models + the remaining two learners agree on the qualitative mixed-regime pattern (QCWA $>$ QWA $\geq$ AO).

### C.2. TruthfulQA: Case-3 robustness across target models and learners

Headline QWA / AO AUPRC at Case 3. LogReg-L2 $\times$ both target models are reproduced from main Table 2.

*Table 6.* TruthfulQA Case-3 (HS-dynamics) cross-target model and cross-learner robustness. Macro AUPRC for QWA/ AO and the QWA−AO gain. QWA exceeds AO across all six (target model, learner) cells.

| Target model | Learner | QWA | AO | gain (QWA−AO) |
|---|---|---|---|---|
| GPT-OSS-20B | LogReg-L2 | 0.664 | 0.641 | +0.023 |
| GPT-OSS-20B | XGBoost | 0.696 | 0.629 | +0.067 |
| GPT-OSS-20B | MLP | 0.682 | 0.638 | +0.044 |
| Qwen3-4B-AWQ | LogReg-L2 | 0.715 | 0.638 | +0.077 |
| Qwen3-4B-AWQ | XGBoost | 0.722 | 0.653 | +0.068 |
| Qwen3-4B-AWQ | MLP | 0.690 | 0.633 | +0.057 |

QWA exceeds AO across all six (target model, learner) cells. The QWA−AO gain is numerically larger for Qwen3-4B-AWQ ($\approx +0.06$–$0.08$) than for GPT-OSS-20B ($\approx +0.02$–$0.07$) in this experiment, though we do not attribute this difference to model size alone—tokenizer, architecture, and AWQ quantization also differ between the two target models.

## D. Dataset Statistics

Per-variant answer length and perplexity. The answer-token count is computed under the corresponding target model's tokenizer. Perplexity is the TF-derived $\exp(-\text{logprob})$.

### D.1. PHANTOM context, answer length, and perplexity

We pre-filter PHANTOM long-context variants to those whose GPT-OSS-20B-tokenized context length lies in 7K–14K tokens, then deduplicate by seed example. The resulting context-length distribution (GPT-OSS-20B tokenizer) is summarized in Table 7. The median $\approx 11.2$K tokens places the experiment in a genuine long-context regime; we report the same long-context subset under both target models (token counts under Qwen3-4B-AWQ tokenizer differ slightly because of vocabulary differences but fall in the same order of magnitude).

*Table 7.* PHANTOM long-context subset: token-length distribution (GPT-OSS-20B tokenizer).

| Quantity | min | p10 | median | mean | p90 | max |
|---|---|---|---|---|---|---|
| Context tokens | 7,040 | 10,304 | 11,200 | 11,263 | 12,863 | 13,794 |
| Total tokens | 7,080 | 10,392 | 11,265 | 11,337 | 12,913 | 13,823 |

*Table 8.* PHANTOM answer length and perplexity by variant.

| Target model | Variant | $n_q$ | tok median | tok mean | tok std | PPL median | PPL mean | PPL std |
|---|---|---|---|---|---|---|---|---|
| GPT-OSS-20B | GT | 396 | 40 | 48.3 | 52.2 | 1.97 | 16.08 | 247.3 |
| GPT-OSS-20B | PHANTOM HN | 396 | 86 | 94.1 | 34.6 | 4.52 | 4.97 | 1.88 |
| Qwen3-4B-AWQ | GT | 396 | 42 | 51.4 | 58.6 | 5.89 | (extreme outliers) | – |
| Qwen3-4B-AWQ | PHANTOM HN | 396 | 92 | 99.1 | 35.6 | 7.58 | 8.90 | 4.86 |

The Qwen3-4B-AWQ GT mean perplexity is dominated by AWQ-quantization extreme outliers (mean $\approx 1.1 \times 10^{15}$, std $\approx 2.3 \times 10^{16}$); we report only the median for this row (see also Appendix A.3). Note the systematic GT/HN length asymmetry on PHANTOM (HN $\approx 2\times$ GT in median tokens) — this is the source of the length confound flagged in Result 2 and Table 1.

### D.2. TruthfulQA answer length

*Table 9.* TruthfulQA answer-token length by variant under each target model's tokenizer. GT and "Incorrect" distributions are essentially identical, removing the gross length confound observed on PHANTOM.

| Target model | Variant | $n_q$ | tok median | tok mean | tok std |
|---|---|---|---|---|---|
| GPT-OSS-20B | GT (best answer) | 817 | 10 | 10.3 | 4.6 |
| GPT-OSS-20B | Incorrect | 817 | 9 | 9.9 | 4.8 |
| Qwen3-4B-AWQ | GT (best answer) | 817 | 10 | 10.5 | 4.7 |
| Qwen3-4B-AWQ | Incorrect | 817 | 10 | 10.1 | 5.0 |

GT and incorrect distributions are essentially identical, largely removing the gross length confound that drives PHANTOM's AO baseline (individual per-query length differences and tokenizer-level effects may still contribute marginally).

## E. Attention Case Studies

### E.1. Successful QCWA context routing

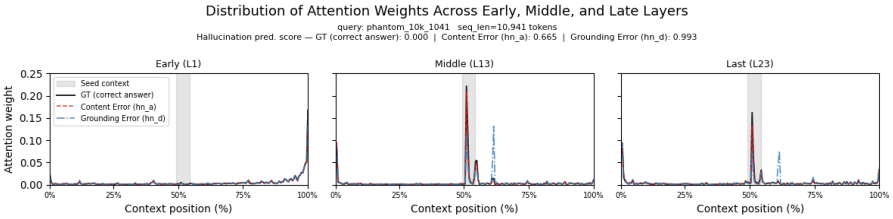


*Figure 2.* PHANTOM 10K, success case: localized attention to seed context.

Figure 2 shows a QCWA run on a PHANTOM 10K example where the probe correctly classifies GT/HN: attention concentrates on the seed-context tokens for GT / faithful HN, and away from the seed context for grounding-error HN. Naively, this looks like context-grounding evidence—the probe "found" the seed and used it to decide. However, the AO controls in Table 1 and Appendix C show that AO alone already recovers $80.6\%$ of the above-baseline QCWA separability on this benchmark; the same probe achieves near-equivalent decisions when the prompt and context tokens are removed at TF time. So even in this individually correct case, the apparent context routing is not the actual basis for the aggregate decision: the AO component dominates.

### E.2. Diffuse-attention failure case

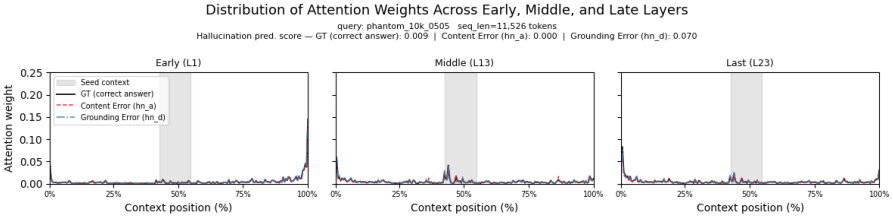


*Figure 3.* PHANTOM 10K, failure case: diffuse attention without clear seed-context routing.

In contrast to Figure 2, Figure 3 shows a QCWA run where the probe fails to reliably separate GT from HN, and the attention distribution does not concentrate on any specific seed-context span. Together, Figures 2 and 3 illustrate that QCWA-side attention patterns are heterogeneous and do not aggregate into a "QCWA recruits context" narrative without the AO/QWA/QCWA controls reported in Tables 1–2.

## F. GT-Conditioned Gemini HNs: Construction Protocol

The GT-conditioned Gemini-HN benchmark used in Appendix B is constructed by synthesising hard-negative answers from existing RAG QA triples. We use as source data the (query, retrieval context, GT-answer) triples of PHANTOM (Ji et al., 2025). For each triple we generate four type-specific hard-negative answers with **Gemini-2.5-flash** (input: the same query, context, and GT answer; output: a single hard-negative candidate per type), giving five answer variants per query (GT plus four HN types). A sixth *abstain* variant is drawn uniformly from a fixed pool of refusal templates and is excluded from CV evaluation because it is trivially separable in pilot experiments.

**Type-specific generation.** Each HN type targets a distinct grounding-failure mode, and the sub-category within a type is chosen automatically by the generator from the (query, context, GT) content:

**hn_a (Content Error).** Fact substitution to a near-by but incorrect value or polarity flip (*contradictory*), or insertion of facts not present in the context (*fabrication*).

**hn_b (Qualification Error).** The surface fact remains correct but important conditions or exceptions are dropped (*critical_omission*), qualifiers are removed yielding overgeneralisation (*overgeneralization*), or false restrictions absent from the context are added (*undue_restriction*).

**hn_c (Structural Error).** Uses *only* facts present in the context but distorts the relations between them: entity/role swap (*entity_role_confusion*), causal/dependency/default-relation distortion (*relationship_distortion*), temporal shift of events or deadlines (*temporal_misinterpretation*), or numerical swap across categories (*numerical_limit_distortion*).

**hn_d (Retrieval/Grounding Error).** Cites facts that do exist in the context but applies them out of scope (*scope_misattribution*), or distorts conditional/independent relations into false dependencies (*evidence_misgrounding*).

**Hardness design.** The construction targets the regime "easy to flag *with* GT, hard to flag *without* GT". HN answers are constrained to match GT in style and length to within $\pm 30\%$, and hn_c / hn_d are restricted to facts that are literally present in the context (only relations or scope are distorted). This is the property the AO audit in Appendix B is designed to stress-test: under the strong content-overlap constraint, an answer-only classifier should *not* be able to discriminate GT from HN unless some answer-side artifact remains.

**Subset used.** The sanity-check experiments in Appendix B run on the long-context Middle-position subset (PHANTOM 10K-token variant; context-length distribution as in Table 7).