Do RAG Hallucination Probes Need the Context?

An Answer-Only Audit

Atsushi Magata (Sompo Japan Insurance) **Makoto Shimomura** (Independent Researcher)

Mechanistic Interpretability Workshop · ICML 2026 · Virtual Poster

Give a RAG hallucination probe the answer alone — no question, no retrieved context — and it still recovers **80.6%** of its full-context separability.

So high activation separability is not, by itself, evidence that a probe is grounding-sensitive.

TL;DR — what the audit shows

1. **Answer-anchored, not grounding.** Fed the answer alone — no question, no context — a probe still recovers 80.6% of a full-context probe's above-chance separability on PHANTOM.

2. **It is dataset-dependent.** PHANTOM is a mixed regime (large answer-only part + a genuine context gain); on TruthfulQA the answer-only signal is weak and the query adds signal. Hash/shuffle controls collapse to chance.

3. **Audit before you trust.** Report AO alongside QCWA and control answer length & style; pre-generation probes that never see the answer sidestep the shortcut entirely.

1 Background & motivation

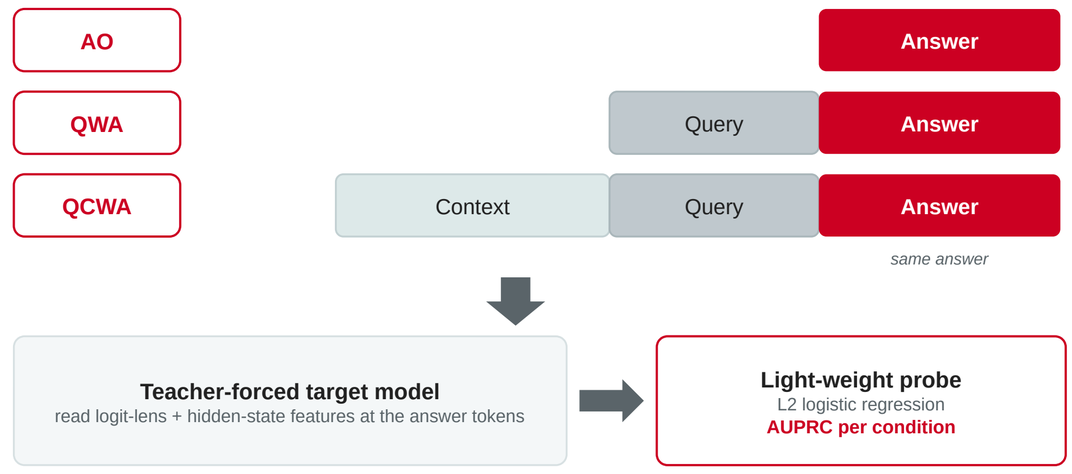
Teacher-forced internal features — logit-lens trajectories and hidden-state dynamics — strongly separate grounded answers from hallucinated hard negatives on RAG benchmarks, and recent work reads hallucination signals directly from LLM hidden states.

But NLP has a recurring failure pattern: probes exploit one-sided artifacts (e.g. hypothesis-only shortcuts in NLI) and length / style cues that outweigh substance. A probe reading activations at the answer tokens could do the same — separating grounded from hallucinated answers without ever using the retrieved context.

So high activation separability is not, by itself, evidence of grounding. Does this failure mode affect mechanistic RAG hallucination detectors? We make it measurable with a simple audit (right).

2 The audit protocol

The SAME candidate answer is fed under three input conditions; we read features at the answer tokens and compare per-condition AUPRC:



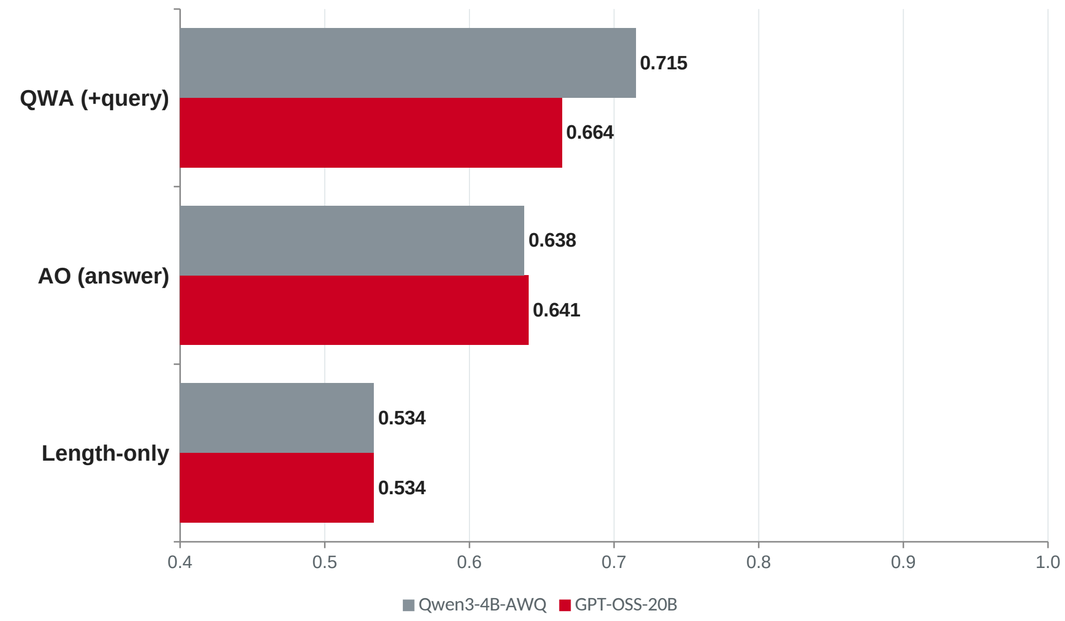
The answer (red) is identical in every condition — we only add the query and context in front, then compare per-condition AUPRC. A high AO score means the signal is answer-anchored, not grounding.

3 Results — the answer alone carries most of the signal

Sanity check. Hash-answer and label-shuffle controls collapse every feature to chance — confirming the signal is the natural answer tokens, not pipeline leakage.

TruthfulQA — weaker answer-only signal

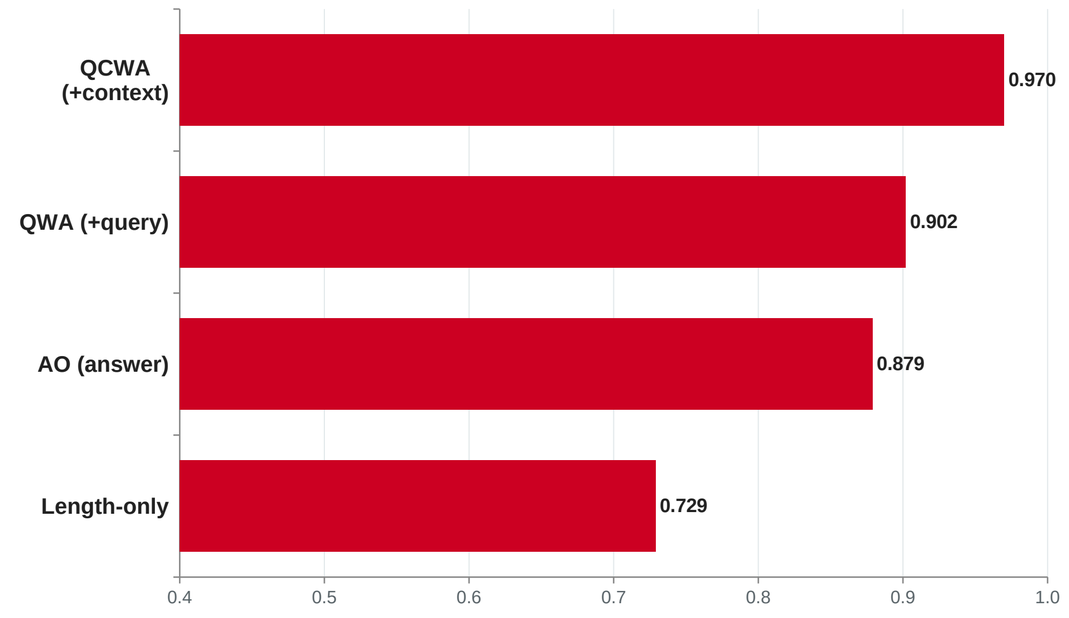
Short, human-authored answers · no retrieval context



Length-only collapses to ≈0.53. **QWA consistently beats AO** (+0.03 / +0.08) — query-conditioned compatibility adds real signal beyond the answer alone.

PHANTOM — a mixed regime

Long-context financial QA · GPT-OSS-20B, Case 3 (HS dynamics)



QCWA (0.970) dominates, yet **AO (0.879) alone recovers 80.6%** of above-chance separability — a genuine context gain sits on top of a large answer-only component.

4 Limitations & scope

We do **not** claim context is unused — PHANTOM shows a real QCWA−AO gain — nor that internal features are uninformative; the claim is interpretive: such features need **AO / QWA / QCWA controls** before being called grounding-sensitive. **Scope:** synthetic generation modes; answer length/style and human-vs-synthetic effects not fully disentangled. Target models: GPT-OSS-20B, Qwen3-4B-AWQ.

Paper & code: github.com/amagata-dx/tf-probe **Contact:** amagata3@sompo-japan.co.jp

PAPER CODE