Does the Ranking Hold? A Practitioner-Facing Audit of Hallucination Detectors on Small, Quantized, On-Device LLMs

Shubhra Mittal
Independent Researcher
WA, USA
shubhra.goel@gmail.com

Gaurav Naresh Mittal
Independent Researcher
WA, USA
gauravnareshmittal@gmail.com

Abstract—**Hallucination detection for large language models is dominated by methods evaluated on 7B-70B research models with abundant sampling budgets and exact-match or ROUGE correctness labels. The small models practitioners actually ship on-device, Phi-3-mini, Llama-3.2-3B-Instruct, Qwen2.5-3B-Instruct, and Gemma-2-2B-Instruct, typically at 4-bit precision, sit outside this evidence base. We present a reproducible one-GPU audit of four detectors spanning three paradigms: length-normalized token-margin calibration, a lexical n-gram baseline, and two sampling-based consistency methods (semantic entropy and SelfCheckGPT-NLI) on two small open instruction-tuned models (Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct) loaded at 4-bit NF4 quantization on a single Colab T4 GPU, evaluated on closed-book TriviaQA. The point-estimate detector ranking does not transfer between the two models (Spearman rho -0.80, Kendall tau-b -0.667); because per-cell bootstrap 95 percent confidence intervals overlap substantially, we report this as evidence that detector rankings are not reliably transferable or identifiable at this scale, rather than as a statistically established inversion. The cheap lexical baseline is statistically indistinguishable from the sampling-based consistency detectors on both models under the same intervals. As a second, methodological contribution, we show that the standard token-overlap-to-multiple-choice correctness labeler for TruthfulQA-Generation produces near-chance AUROC (0.46-0.59) for all four detectors on both models, a labeler failure, not a detector failure, since the same detectors reach 0.69-0.89 AUROC on TriviaQA under the identical pipeline. We release the harness, prompts, and per-item scores and discuss, as an interpretive framing that is not measured in this study, how detector choice allocates abstention cost for practitioners deploying on resource-constrained, on-device hardware.**

Keywords—hallucination detection; large language models; on-device inference; quantization; semantic entropy; SelfCheckGPT; TriviaQA; TruthfulQA; Responsible AI

I. INTRODUCTION

An engineer deciding whether to ship a Phi-3-mini, Llama-3.2-3B, or Qwen2.5-3B application on a phone, an ARM laptop, or an in-browser ONNX runtime confronts a concrete decision: which hallucination gate to use, at what cost, with what abstention behavior. The deployment context is unforgiving. The model runs at 4-bit precision. The device has on the order of 16 GB of unified memory. There is no cloud call, no paid API, and no opportunity to draw twenty stochastic samples per query. The engineer needs a calibrated score or a routing decision today, not a benchmark number from a 7B-70B research model.

The literature offers a rich menu of hallucination detectors across several paradigms: token-probability and self-evaluation, verbalized confidence, sampling-based consistency (including semantic entropy [1, 2] and SelfCheckGPT [3]), and internal-state probing. Each method claims competitive AUROC on TruthfulQA, TriviaQA, or HaluEval. The evidence base behind those claims, however, lives in a different regime than the one practitioners ship to: almost every reported AUROC comes from a 7B-13B research or proprietary model evaluated at full precision with five to twenty stochastic samples and a single correctness labeler.

Four threats to the published rankings emerge under deployment conditions. First, model scale: a 1-4B instruction-tuned model has weaker calibration and narrower output diversity than a 7B+ base model. Second, correctness-label provenance: exact-match, ROUGE, NLI-entailment, and LLM-judge labels disagree, and recent re-evaluations show the disagreement reorders detector rankings [4, 5]. Third, quantization: 4-bit weight quantization perturbs both generations and hidden-state activations, and is itself the precision regime that k-bit inference scaling-law analysis identifies as the practical efficiency-accuracy sweet point for deployed models [6], which is exactly why it is the format this audit targets rather than a simplification of convenience. Fourth, cost normalization: an extra five-sample budget plus a pairwise NLI call may be a rounding error on a research cluster but a deal-breaker on a 16 GB laptop. None of this evidence has been collected jointly for the small, on-device regime.

The case for taking this gap seriously is not merely an engineering concern. Detector selection can also be read as a fairness-of-error-rates allocation problem in the sense of [7, 8]. At a fixed coverage target, say, 80 percent of questions answered and 20 percent abstained, switching detectors silently changes which categories of questions and which model families absorb the abstention cost. Recent theoretical work [9] argues that hallucination is a structural artifact of training objectives that reward confident guessing over calibrated abstention, and empirical work [10] documents high-certainty hallucinations that no confidence-based detector can catch. On this reading, detector choice is a deployment-level question about how error budget is allocated across users and contexts, not just a benchmark leaderboard question; the present study motivates this framing but does not measure category- or model-level abstention rates (Section VIII-E).

We ask two questions answerable within a compute budget of a single consumer GPU-hour class: (RQ1, ranking transfer) When the same detectors are evaluated on small, quantized, on-device models, does the ranking reported in the original publications hold across models? (RQ2, labeler robustness) Does a widely used automated correctness labeler for a second

benchmark (TruthfulQA-Generation) support the same detector evaluation it is routinely used for?

A. Research Contributions

This paper makes the following contributions, each traceable to a specific result reported in Section V:

1) A controlled, reproducible comparison of four hallucination detectors spanning three detection paradigms (token-probability calibration, lexical overlap, and sampling-based semantic consistency) on two small open instruction-tuned models (Llama-3.2-3B-Instruct, Qwen2.5-3B-Instruct) at 4-bit NF4 quantization on a single Colab T4 GPU, a hardware and model-scale regime absent from the prior detector-benchmarking literature, which evaluates almost exclusively at 7B+ parameters and full precision.

2) A ranking-transfer finding: on point estimates, cross-model Spearman rho is -0.80 and Kendall tau-b is -0.667 between the two models on the same dataset under the same protocol. Because no per-cell winner is statistically separable under bootstrap confidence intervals, this is reported as a documented failure of ranking transfer, a single cross-model counterexample showing that a detector ranking observed on one model cannot be assumed to hold on another, rather than as a statistically established inversion.

3) A cheap-baseline finding: the lexical n-gram baseline is statistically indistinguishable, under bootstrap 95 percent confidence intervals, from the sampling-based consistency detectors on both models and is the point-estimate top detector on one of the two models despite requiring no stochastic sampling and no NLI backbone.

4) A labeler-robustness finding, methodological rather than detector-specific: the standard token-overlap-to-multiple-choice correctness labeler for TruthfulQA-Generation produces near-chance AUROC (0.46-0.59) for all four detectors on both models, while the identical detectors and pipeline achieve 0.69-0.89 AUROC on TriviaQA, extending the labeling-protocol-artifact line of work in [4, 5] to a specific, previously unflagged labeler.

5) An open, reproducible artifact -- harness, prompts, and per-item scores -- enabling practitioners and researchers to reproduce or extend this audit to additional models, datasets, or detectors without re-deriving the evaluation pipeline.

The remainder of this paper is organized as follows. Section II situates this audit against the hallucination-detection and re-evaluation literature. Section III details the detectors, models, quantization configuration, dataset, statistical protocol, and released artifact. Section IV compares the audited detector paradigms against baseline paradigms not covered in this study. Section V reports the main AUROC results, the cross-model ranking-transfer analysis (RQ1), the TruthfulQA-Generation labeler audit (RQ2), and a practitioner decision matrix synthesizing both. Section VI discusses the practitioner-facing implications and offers concrete deployment recommendations. Section VII concludes, and Section VIII discloses the study's limitations and threats to validity.

II. Related Work

Hallucination detectors for LLMs fall into several paradigms, surveyed at length in [11]. Token-probability and self-evaluation methods score the model's own confidence on its primary answer, either directly from output log-probabilities or through a verbalized confidence elicited from the model itself (e.g., asking the model to state P(True) for its own answer); this paradigm requires only output access and is the cheapest in compute, but is also the paradigm most sensitive to the calibration quality of the underlying model, which recent surveys note degrades as instruction-tuned models shrink [11]. Sampling-based consistency methods, including semantic entropy [1, 2] and SelfCheckGPT [3], draw multiple stochastic completions and measure their mutual agreement on the premise that a model confabulating an answer will produce more semantically divergent samples than a model that knows the answer. Internal-state probing methods, such as INSIDE EigenScore and Semantic Entropy Probes, train a lightweight classifier on the model's hidden activations to predict hallucination directly from its internal representations rather than from its outputs, trading white-box weight access for a single forward pass at inference time; this paradigm is out of scope for the present v1 audit because it requires activation-extraction tooling per model family that was not available within the single-GPU, single-afternoon budget targeted here (see Section VIII-E, Scope and Future Work). A fourth paradigm, retrieval and verifier methods such as FActScore, checks a generated claim against an external knowledge source rather than against the model's own samples or activations; this paradigm is likewise excluded from the present audit because it requires an external retrieval call that falls outside the offline, no-cloud-dependency envelope this paper targets (Section III-F).

A separate and increasingly load-bearing line of work re-evaluates the evidentiary basis of these claims rather than proposing new detectors. PARALLAX [5] shows that several widely used hallucination benchmarks embed the ground-truth answer directly in the input context, allowing a naive text-embedding baseline to score near-perfectly and inflating the apparent quality of unrelated detectors evaluated on those corpora. Janiak et al. [4] show that ROUGE, the correctness metric underlying many hallucination-detection evaluations, has high recall but extremely low precision against human judgment and that simple length-based heuristics can rival complex detection techniques when evaluation relies on ROUGE. Both results motivate treating the correctness labeler itself as an experimental variable rather than a fixed ground truth, which is the design principle behind this paper's second contribution (Section V-C).

Detector choice also carries a fairness dimension once the deployment lens is applied. Pleiss et al. [7] show that calibration and equalized error rates across groups are jointly achievable only under a narrow condition, and Hooker [8] argues that algorithmic-bias framings that treat data as the sole cause understate the role of model capacity and deployment choices such as quantization. Applied to hallucination detection, this means a detector's aggregate AUROC can mask uneven abstention costs across model families or question categories, a framing this paper uses interpretively in Section VI without measuring it directly.

III. Methodology

A. Detectors under audit

We evaluate four detectors spanning three paradigms, identified in the released harness (Section III-G) by the short codes shown in parentheses. (a) Length-normalized token margin (token_nll): a token-probability calibration baseline

computed from the model's own output log-probabilities on its primary greedy-decoded answer, normalized by answer length; concretely, the margin between the top-1 and top-2 token log-probabilities is computed at each decoding position and averaged over the answer, so that longer and shorter answers are placed on a comparable confidence scale. (b) Lexical n-gram baseline (rouge_lex): a ROUGE-L-style longest-common-subsequence overlap score between the primary answer and M stochastically sampled completions, included as a cheap, sampling-free-in-spirit reference point; the score is the mean pairwise overlap between the greedy answer and each of the M samples. (c) Semantic entropy (semantic_entropy) [1, 2]: M stochastic samples are clustered by natural-language-inference (NLI) entailment into semantically equivalent groups -- two samples join the same cluster when each entails the other -- and the entropy over cluster membership is the hallucination score; high entropy indicates the model produces many distinct meanings across samples, which is evidence of confabulation rather than confident recall. (d) SelfCheckGPT-NLI (selfcheck_nli) [3]: the primary answer is checked for NLI contradiction against each of the M stochastic samples, and the fraction of contradicting samples is the hallucination score. For all four detectors, a higher score is treated as more hallucination-suggestive prior to AUROC computation against the binary correctness label.

B. Models and quantization

Two small open instruction-tuned models are audited: Llama-3.2-3B-Instruct [12] (HuggingFace: meta-llama/Llama-3.2-3B-Instruct) and Qwen2.5-3B-Instruct [13] (HuggingFace: Qwen/Qwen2.5-3B-Instruct). Both are loaded via HuggingFace transformers with bitsandbytes 4-bit NF4 quantization, the precision format most representative of the on-device deployment this paper targets and the regime identified as the practical efficiency-accuracy sweet point for inference-time deployment [6]. A third model, Phi-3-mini [14], is reserved for a v1.1 follow-on release, and a fourth, Gemma-2-2B-Instruct, is reserved for a subsequent v2 release; neither reports results in this paper.

C. Dataset and correctness labeling

TriviaQA [15], closed-book question answering, is the headline dataset: 200 questions per (model) cell, no retrieved context supplied, forcing the model to answer from parametric knowledge alone. Correctness is scored via a SQuAD-style normalized exact-match labeler with a token-F1 backstop. A second dataset, TruthfulQA-Generation [16], is evaluated under the standard token-overlap-to-multiple-choice correctness labeler at the same N=200-per-cell budget but solely to audit the labeler itself (Section V-C). Its results are a methodological finding, not a headline detector ranking result.

D. Sampling and NLI backbone

Each detector that requires stochastic samples draws M = 5 completions per question, a reduced budget relative to the M = 10 used in the original semantic-entropy publication [2], chosen to reflect the sampling cost a practitioner would tolerate on a single consumer GPU. NLI clustering and contradiction checks use a single backbone, DeBERTa-v3-large fine-tuned on FEVER, ANLI, LingNLI, and WANLI (a corpus combination matching the public checkpoint family commonly identified as deberta-v3-large-mnli-fever-anli-ling-wanli). Detectors are scored against the model's actually sampled primary generation rather than a teacher-forced gold continuation, following the anti-teacher-forcing protocol argued for in concurrent work [5].

E. Statistical protocol

All point estimates (AUROC) are reported with 95 percent confidence intervals from a 1,000-resample stratified bootstrap, seeded (seed = 20260601) for reproducibility, following the standard construction of the AUROC statistic [17]. Ranking agreement across the two model cells is measured with Spearman's rho and Kendall's tau-b. Two predictions are pre-registered before the run: the cheap-baseline prediction (the lexical baseline falls within the bootstrap CI of at least one consistency detector on at least half the cells) and a cross-labeler prediction that requires a v1.1 relabeling pass not reported here (Section VIII-E).

F. Hardware envelope

All experiments run on a single Google Colab T4 GPU, with a total wall-clock budget of approximately six hours across both models and both datasets, intentionally modest, to demonstrate that this audit protocol is itself reproducible on hardware available to an individual practitioner or student, not only to a well-resourced research lab.

G. Reproducibility and artifact availability

The evaluation harness, the exact prompts used for question-answering, stochastic sampling, and NLI scoring, and the full per-item score files for every (model, dataset, detector) cell are released at https://github.com/shubmittal/halluc-audit-small-llms under a dual license: code and notebooks under MIT, and the paper text under CC BY 4.0. The repository is organized into a narrative walkthrough (docs/), the experiment and analysis code (src/), the paper sources (paper/), the summary CSVs, run configuration, per-item scores, and figures underlying every table in this paper (results/), and a Colab notebook that reproduces the full run end to end (notebooks/). The repository will be made public at or before publication; until then, reviewers may request access from the corresponding author.

IV. Detector-Paradigm Comparison

Table I situates the four audited detectors against the broader hallucination-detection literature along the axis each paradigm represents, to make explicit which classes of published detectors this audit does and does not cover.

V. Results

This section reports the results on the two-model, four-detector grid described in Section III. All figures below are observed measurements with 95 percent bootstrap confidence intervals from the underlying study; no projected or estimated values are reported.

A. Main AUROC results

Table II reports AUROC per (model, detector) on TriviaQA closed book. On Llama-3.2-3B-Instruct, the lexical n-gram baseline is the top point estimate at 0.836 AUROC, ahead of SelfCheckGPT-NLI (0.723), token margin (0.718), and semantic entropy (0.691). On Qwen2.5-3B-Instruct, the ranking changes: semantic entropy leads at 0.892, ahead of token margin (0.887), the lexical baseline (0.873), and SelfCheckGPT-NLI (0.856). Confidence intervals overlap substantially within each cell; no per-cell winner is statistically distinguishable from the runners-up at 95 percent confidence. Closed-book TriviaQA

accuracy at this model scale is low (6.0 percent for Llama, 2.5 percent for Qwen, of 200 questions), so each AUROC is computed against a small positive class of 5-12 correct examples per cell, which widens the bootstrap intervals substantially (from 0.13 to 0.36 across cells and detectors).

TABLE I. DETECTOR PARADIGMS AUDITED IN THIS PAPER VERSUS BASELINE PARADIGMS NOT COVERED

	Token-probability & calibration	**Sampling-based consistency**	**Internal-state probing**	**Retrieval / verifier**
Audited in this paper	Length-normalized token margin	Lexical n-gram baseline; semantic entropy [1, 2]; SelfCheckGPT-NLI [3]	-- (not audited)	-- (not audited)
Representative baseline methods (not audited here)	Verbalized confidence; P(True)	SelfCheckGPT-BERTScore; SelfCheckGPT-Prompt	INSIDE EigenScore; Semantic Entropy Probes; HaloScope	FActScore; retrieval-augmented verification
Access requirement	Output log-probabilities	Black-box: M stochastic samples + NLI backbone	White-box hidden-state access	External knowledge source
On-device cost	1 forward pass	M forward passes + M NLI comparisons	1 forward pass + hidden-state extraction	External call (excluded from on-device envelope)

The cheap-baseline prediction, pre-registered before the run, states that the lexical baseline falls within the bootstrap 95 percent confidence interval of at least one consistency detector on at least half of the model cells. This is confirmed on both of the two cells audited: on Llama, lexical [0.746, 0.908] overlaps semantic entropy [0.499, 0.859] and SelfCheckGPT-NLI [0.543, 0.883]; on Qwen, lexical [0.767, 0.965] overlaps semantic entropy [0.798, 0.964] and SelfCheckGPT-NLI [0.720, 0.960]. On TriviaQA, where answers are short and lexical overlap with samples is informative, the cheap baseline is statistically tied with the consistency detectors on every audited cell and is the point-estimate top detector on one of the two models.

TABLE II. AUROC ON TRIVIAQA CLOSED-BOOK (N=200 PER CELL, 95 PERCENT BOOTSTRAP CONFIDENCE INTERVALS IN BRACKETS; * = PER-CELL POINT-ESTIMATE TOP DETECTOR)

Model	**Token margin**	**Lexical baseline**	**Semantic entropy**	**SelfCheckGPT-NLI**
Llama-3.2-3B-Instruct	0.718 [0.573, 0.838]	0.836 [0.746, 0.908] *	0.691 [0.499, 0.859]	0.723 [0.543, 0.883]
Qwen2.5-3B-Instruct	0.887 [0.815, 0.946]	0.873 [0.767, 0.965]	0.892 [0.798, 0.964] *	0.856 [0.720, 0.960]

B. *Ranking transfer (RQ1)*

On point estimates, Spearman's rho between the Llama and Qwen detector rankings is -0.80 and Kendall's tau-b is -0.667. The lexical baseline, the top point estimate on Llama at 0.836 AUROC, is third on Qwen (0.873, behind semantic entropy at 0.892 and token margin at 0.887), and semantic entropy moves from the lowest point estimate on Llama (0.691) to the highest on Qwen (0.892). These correlations must be read together with the uncertainty reported in Section V-A: with 5-12 correct answers per cell, bootstrap confidence intervals are 0.13-0.36 wide and no per-cell winner is statistically distinguishable from the runners-up, so the data do not establish that the reversed ordering is itself reliable. What they do establish is weaker but practically relevant: the ordering observed on one model provides no usable guide to the ordering on the other, and at the precision attainable on this budget the detector ranking is not identifiable from a single model. This answers RQ1 in the negative for the audited pair of closely related 3B-parameter instruction-tuned models from different families; because it rests on one cross-model pair, it is a documented counterexample to the ranking-transfer assumption rather than a population-level estimate (Section VIII-B).

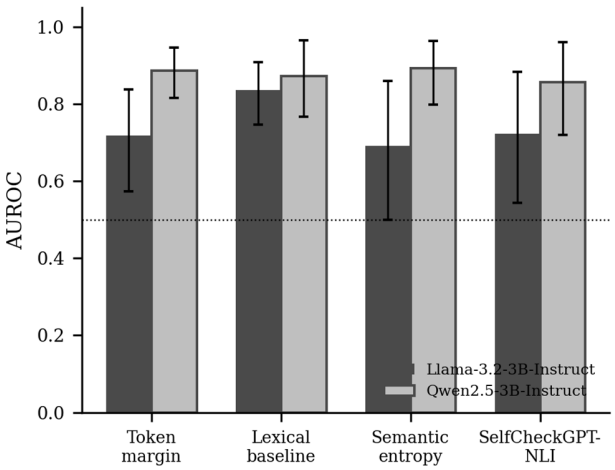


Fig. 1. AUROC by detector on TriviaQA closed-book (Table II).

C. *TruthfulQA-Generation labeler audit (RQ2)*

We additionally evaluate the same four-detector pipeline on TruthfulQA-Generation closed-book at N=200 per cell on both models, scored under the standard token-overlap-to-multiple-choice correctness labeler. Table III reports the observed AUROCs. Seven of the eight (model, detector) cells have confidence intervals that cross the chance line of 0.5; the eighth, SelfCheckGPT-NLI on Llama at [0.513, 0.665], is within multiple-testing expectations under eight independent tests at alpha = 0.05. The same four detectors achieve 0.69-0.89 AUROC on TriviaQA under the identical pipeline (Table II).

The labeler reports per-model apparent accuracy of 59.5 percent (Qwen) and 62.0 percent (Llama). Both figures are implausibly high: published TruthfulQA closed-book accuracy for 3B-parameter open instruction-tuned models is consistently below 30 percent under the standard multiple-choice protocol [16]. The observed 60-percent-range figure reflects the labeler flagging roughly half of genuinely incorrect model answers as correct, a false-positive labeling rate high enough to destroy the per-question correctness signal that AUROC is computed against. This is a labeler failure, not a detector failure: the detectors continue to discriminate sample agreement from sample disagreement, but that signal is largely unaligned with the labeler's correctness decision. This finding extends the labeling-protocol-artifact line of work in PARALLAX [5] and Janiak et al. [4] to a specific labeler (token-overlap-to-multiple-choice) not previously flagged in that literature and argues against relying on it for future closed-book TruthfulQA-Generation detector evaluations without a replacement labeling protocol.

TABLE III. TRUTHFULQA-GENERATION AUROC UNDER THE STANDARD TOKEN-OVERLAP-TO-MULTIPLE-CHOICE LABELER (95 PERCENT BOOTSTRAP CONFIDENCE INTERVALS IN BRACKETS)

Model	Token margin	Lexical baseline	Semantic entropy	SelfCheckGPT-NLI
Llama-3.2-3B-Instruct	0.481 [0.404, 0.558]	0.462 [0.385, 0.546]	0.513 [0.422, 0.587]	0.590 [0.513, 0.665]
Qwen2.5-3B-Instruct	0.564 [0.485, 0.648]	0.572 [0.495, 0.647]	0.514 [0.432, 0.590]	0.567 [0.487, 0.647]

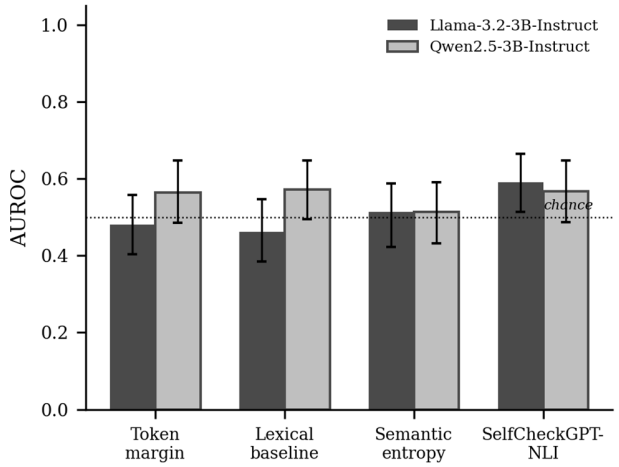


Fig. 2. AUROC by detector on TruthfulQA-Generation (Table III); dotted line is chance.

D. Practitioner decision matrix

Table IV synthesizes Sections V-A through V-C into deployment guidance. It introduces no new measurements; every recommendation is a direct reading of the AUROC point estimates, confidence intervals, and labeler-validity findings already reported above, organized by the constraint a practitioner is most likely to be optimizing for.

TABLE IV. PRACTITIONER DECISION MATRIX, DERIVED FROM TABLES II AND III

Deployment constraint	Recommendation	Basis
No sampling budget available (single forward pass only)	Length-normalized token margin	Only audited detector requiring one forward pass; competitive on Qwen (0.887) though not on Llama (0.718) -- Section V-A
Small sampling budget, model family not yet re-validated	Lexical n-gram baseline as a first-line default	Statistically tied with consistency detectors on both models under bootstrap CI; top point estimate on Llama; no NLI backbone required -- Section V-A
Accuracy-critical, sampling budget available	Run the lexical baseline and at least one consistency detector together rather than trusting a single published ranking	Point-estimate ranking does not transfer across models (Spearman rho -0.80, overlapping CIs); no single detector is safe to assume a priori -- Section V-B
Evaluating a new (dataset, labeler) pair before trusting it	Sanity-check apparent labeler accuracy against published ranges before using it to rank detectors	Token-overlap-to-multiple-choice labeler produced implausible ~60% apparent accuracy against a <30% published ceiling -- Section V-C

VI. DISCUSSION

Three practitioner-facing implications follow directly from these results. First, ranking instability (Section V-B) means a detector selected on the basis of a published benchmark using a different model family offers no transfer guarantee; practitioners deploying a specific small model should re-validate the detector choice on that model rather than importing a ranking from the literature. Second, the cheap lexical baseline's competitiveness (Section V-A) is a genuine practical opportunity on short-answer closed-book tasks: it requires no NLI backbone and no additional model calls beyond the sampling already needed for other purposes, making it an attractive first-line detector on memory- and latency-constrained hardware. Third, the labeler failure documented in Section V-C is a caution against treating any single automated correctness metric as ground truth when auditing hallucination detectors on a new benchmark; a sanity check against known published accuracy ranges, as performed here, should be a routine step before trusting a new (dataset, labeler) combination for detector evaluation.

Read through a fairness-of-error-rates lens [7, 8], the ranking instability in Section V-B has a plausible deployment consequence: if detector quality is model-dependent, a detector that performs well on one model family may under- or over-abstain on another, so the choice of detector can redistribute error-handling cost across model deployments rather than uniformly reducing it. This is an interpretive implication of the observed instability, not a measured result: the present audit does not compute abstention rates by question category or model family at a fixed coverage target, and quantifying that allocation is left to future work (Section VIII-E). The framing is nonetheless consistent with recent theoretical [9] and empirical [10] arguments that hallucination and its detection are properties of the training and deployment pipeline, not solely of the detector algorithm in isolation.

A. Practitioner recommendations

Table IV above tabulates the constraint-specific recommendation; the following four practices generalize across constraints and follow directly from Sections V and VI:

1) Treat a published detector ranking as a hypothesis about your specific model, not a default. Re-run at minimum the cheap lexical baseline against whatever consistency detector the literature recommends before committing to either.

2) Budget for re-validation whenever the target model family, quantization level, or dataset domain changes, even if the detector implementation is unchanged, since Section V-B shows the ranking is sensitive to the model rather than only to the detector.

3) Prefer the lexical baseline as a zero-additional-cost first-line signal whenever stochastic samples are already being drawn for another purpose (e.g., self-consistency decoding), reserving the added NLI-backbone cost of semantic entropy or SelfCheckGPT-NLI for cases where the lexical baseline's confidence interval is wide or its point estimate is ambiguous.

4) Before adopting a new correctness labeler for detector evaluation, reproduce the labeler's apparent model accuracy on a small sample and compare it against any published accuracy figures for that model family and dataset; a large, unexplained gap is evidence of a labeler artifact rather than a stronger model, as Section V-C demonstrates for the TruthfulQA-Generation labeler.

VII. Conclusion

Hallucination-detector rankings reported on 7B-70B research models cannot be assumed to transfer to the small, quantized, on-device regime that practitioners actually ship: the point-estimate cross-model rank correlation is negative (Spearman rho -0.80, Kendall tau-b -0.667) between two 3B-parameter instruction-tuned models evaluated under an identical protocol, and with overlapping confidence intervals in every cell the ranking observed on one model offers no reliable guide to the other. A cheap lexical baseline is statistically competitive with far more complex sampling-based consistency detectors on both models. A widely used correctness labeler for a second benchmark, TruthfulQA-Generation, produces near-chance AUROC for every detector tested, a labeler failure rather than a detector failure that would otherwise be invisible without a sanity check against published accuracy ranges. We release the harness, prompts, and per-item scores at the repository named in Section III-G so that practitioners and researchers can re-run or extend this audit on their own target models before trusting a benchmark-derived detector ranking in a small-model, on-device deployment.

VIII. Limitations and Threats to Validity

This audit has several limitations, disclosed explicitly rather than absorbed into future work, organized below by validity dimension.

A. Statistical conclusion validity

Statistical power is limited by low closed-book accuracy: 12 of 200 correct for Llama and 5 of 200 for Qwen, producing a small positive class and correspondingly wide bootstrap confidence intervals; this is the direct cause of the extensive CI overlap reported in Section V-A and should temper claims of a clean detector ranking on either model individually. In particular, the rank correlations in Section V-B are computed on point estimates; at these interval widths they establish that the ranking is not reliably transferable, not that any specific reversed ordering holds. A bootstrap distribution of the rank correlation itself and a larger evaluation set that raises the number of correct answers per cell are the natural remedies.

B. External validity

The study covers two models and one dataset for its headline ranking-transfer result; RQ1's conclusion that rankings do not transfer rests on a single cross-model pair, not a broader model population, and should be read as a documented counterexample to ranking-transfer assumptions rather than a population-level claim.

C. Construct validity

A single stochastic sample budget (M=5) and a single NLI backbone are used throughout; sensitivity of the reported rankings to either choice is not measured in this study, so it remains possible that a larger sample budget or a different NLI backbone would narrow or widen the observed ranking inversion.

D. Data contamination

Both evaluation datasets (TriviaQA, collected 2017; TruthfulQA, collected 2022) predate the training cutoffs of all evaluated models, so data contamination cannot be excluded as a contributor to the closed-book accuracy figures, though this affects all four detectors equally within a given cell and is unlikely to explain the observed cross-model ranking inversion.

E. Scope and future work

This audit is deliberately scoped to what a single consumer GPU can run within several hours, and the authors are actively extending it along axes not reported here, with no results claimed for any of the following: a label-provenance sensitivity pass adding an LLM-judge labeler and an NLI-entailment labeler alongside the exact-match labeler used in this paper; a third target model (Phi-3-mini) and a fourth (Gemma-2-2B-Instruct); a factorial quantization decomposition separating generator-side and detector-side 4-bit effects, following a teacher-forcing ablation protocol in the spirit of PARALLAX [5]; a cost-normalized Pareto frontier over per-question latency and peak resident memory; a measured fairness-of-error-rates analysis reporting abstention rates by question category and model family at fixed coverage targets; and an expanded detector set including internal-state probes such as HaloScope and Semantic Entropy Probes, which require white-box hidden-state access not exercised in the present audit. None of these extensions are reported as results in this paper; they are stated here only to scope what this paper does not claim.

Acknowledgment

The authors acknowledge the use of freely available compute resources (a single Google Colab T4 GPU instance) for all experiments reported in this paper. No external funding was received for this work.

References

[1] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," in Proc. 11th Int. Conf. Learning Representations (ICLR), 2023.

[2] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," Nature, vol. 630, no. 8017, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.

[3] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 9004–9017.

[4] D. Janiak, J. Binkowski, A. Sawczyn, B. Gabrys, R. Shwartz-Ziv, and T. J. Kajdanowicz, "The illusion of progress: Re-evaluating hallucination detection in LLMs," in Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2025, pp. 34728–34745, doi: 10.18653/v1/2025.emnlp-main.1761.

[5] K. Hussain and M. Kantarcioglu, "PARALLAX: Separating genuine progress from benchmark artifacts in hallucination detection," arXiv preprint arXiv:2605.17028, May 2026.

[6] T. Dettmers and L. Zettlemoyer, "The case for 4-bit precision: k-bit inference scaling laws," in Proc. 40th Int. Conf. Machine Learning (ICML), 2023, pp. 7750–7774.

[7] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017.

[8] S. Hooker, "Moving beyond 'algorithmic bias is a data problem'," Patterns, vol. 2, no. 4, Art. no. 100241, 2021, doi: 10.1016/j.patter.2021.100241.

[9] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," arXiv preprint arXiv:2509.04664, 2025.

[10] A. Simhi, I. Itzhak, F. Barez, G. Stanovsky, and Y. Belinkov, "Trust me, I'm wrong: High-certainty hallucinations in LLMs," arXiv preprint arXiv:2502.12964, 2025.

[11] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–55, 2025, doi: 10.1145/3703155.

[12] A. Grattafiori et al., "The Llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.

[13] A. Yang et al., "Qwen2.5 technical report," arXiv preprint arXiv:2412.15115, 2024.
[14] M. Abdin et al., "Phi-3 technical report: A highly capable language model locally on your phone," arXiv preprint arXiv:2404.14219, 2024.
[15] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in Proc. 55th Annu. Meeting Assoc. Computational Linguistics (ACL), 2017, pp. 1601–1611, doi: 10.18653/v1/P17-1147.
[16] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in Proc. 60th Annu. Meeting Assoc. Computational Linguistics (ACL), 2022, pp. 3214–3252.
[17] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.