Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Semantic Entropy Hallucination Detection: An Evidence Audit

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-11-02 · semantic entropy hallucination detection · semantic entropy · llm hallucination detection · uncertainty quantification llm · semantic entropy limitations · semantic entropy probes · rag grounding verification · unsupported claims detection · evidence-conditioned evaluation

Original article: https://reliablecontext.com/articles/semantic-entropy-hallucination-detection-evidence-audit



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes in the Semantic Uncertainty Literature, 2023 to 2026
4. Where Sampling Agreement Stops: Consistency, Correctness and Support
5. Data Analysis and Evidence
6. Implementation Considerations and Process Changes
7. Implications and Future Directions
8. Frequently Asked Questions (FAQs)
9. Conclusion

Executive Summary

Semantic entropy (SE) is an uncertainty score computed over the meanings of several sampled answers. It is not a truth test and it is not a source-support test. Farquhar et al. published the Nature method on 19 June 2024 (www.nature.com) and scoped it to confabulations, which they define as "arbitrary and incorrect generations" (www.nature.com). The authors state that the method "does not guarantee factuality because it does not help when LLM outputs are systematically bad" (www.nature.com). Its headline result, an **average AUROC of 0.790** across 30 task and model combinations, is a ranking statistic for predicting wrong answers and not an accuracy (www.nature.com). The paper reports raw averages without standard errors (www.nature.com).

For a source-grounded answer the consequence is direct. If every sampled answer lands in one meaning cluster, the empirical cluster entropy is zero, whether the cited source supports that answer or contradicts it. That is synthetic arithmetic, not a benchmark result, but a September 2026 preprint (version 3, revised 7 September 2026) (arxiv.org) describes the same mechanism: standard SE "becomes zero when all samples form one semantic cluster, even if they are incorrect" (arxiv.org), and the share of hallucinated queries in a single cluster ranges from **39% to 99%** across its datasets (arxiv.org). That preprint studies the setting where no trusted context or reference document is available (arxiv.org), so it does not test source-grounded answers. A 2025 study of retrieval-augmented generation (RAG) found that standard uncertainty estimators, SE included, cannot reliably estimate response correctness there (aclanthology.org).

Two further distinctions matter. Support by a supplied source is a different object from factual correctness: the attribution literature deliberately avoids factuality judgments (arxiv.org), and FActScore ties truthfulness to a particular knowledge source (aclanthology.org). And finite samples can miss meanings entirely: the March 2026 Evidential Semantic Entropy paper argues that a heavier tail of unobserved answers indicates higher overall uncertainty (aclanthology.org).

The decision this report supports is whether a sampling-based detector should trigger review of a source-grounded answer. The evidence reviewed here supports using SE, where its access and sampling costs are acceptable, as a **review-routing signal** whose operating point is chosen on calibration data and frozen before held-out evaluation. It does not support skipping independent source-support review for low-entropy answers. The proposed design is a paired, evidence-conditioned audit: source conditions (sufficient support, decisive

span removed, counterevidence, irrelevant context, legitimate multiple answers) crossed with sampled cluster counts, with truth, support and citation resolution adjudicated separately. **No experiment has been run for this report.** Every number labelled synthetic is arithmetic, and every number attributed to a paper is that paper's result on its own datasets.

Introduction and Background

Practitioners who put language-model answers in front of reviewers face a triage problem. Review capacity is finite, so a detector that ranks answers by risk is attractive. Semantic entropy is the best-known sampling-based candidate. It is also frequently described in terms that outrun its definition: "detects hallucinations", "flags wrong answers", "measures confidence". This report separates four objects that those phrases blur.

- **Variability of sampled meanings.** How spread the model's sampled answers are across distinct meanings, after clustering. This is what SE estimates.
- **Correctness against an independently established fact.** Whether the answer is true, judged by a reference that does not depend on the supplied source.
- **Support by the supplied source snapshot.** Whether the answer's claims follow from the identified source version the system was given. Attribution research frames this as verifying generated text against an independent, provided source (direct.mit.edu).
- **Citation resolvability.** Whether each cited locator, such as a URL plus a quoted passage, still resolves to text that contains the claim at the time of review.

The four can disagree. An answer can be true yet unsupported by the supplied source. An answer can be supported by a source that is itself wrong, and an answer can be true yet absent from the source it cites. The NIST Generative AI Profile defines confabulation as the production of confidently stated but erroneous or false content (nvlpubs.nist.gov) and adds that confabulations include outputs that diverge from the prompts or other input (nvlpubs.nist.gov). Even that definition leaves open which reference, the world or the input, is the yardstick, which is why the audit below records both. NIST also notes that generated outputs may include confabulated logic or citations that purport to justify or explain the answer (nvlpubs.nist.gov), which is the failure that citation resolvability is meant to catch.

Information systems that must defend a value to a reviewer care about the last two objects most, because they tie an answer to a source identity and an observation time. The question here is narrower than whether a detector works in general. It is what a detector's score licenses a team to conclude about one answer built on one source version, and what it does not.

The scope is information and evaluation. This report does not rely on any vendor claim, reports no result from a new experiment, and treats every transfer from the cited benchmarks to enterprise source bundles as untested. Dates and version states are as of 2 November 2026.

Key Changes in the Semantic Uncertainty Literature, 2023 to 2026

From token-sequence entropy to meaning-level entropy (2023)

Kuhn, Gal and Farquhar introduced semantic entropy as an entropy over meanings rather than token sequences, one that incorporates linguistic invariances created by shared meanings (iclr.cc). The method is unsupervised, uses a single model and needs no modification of the underlying language model (iclr.cc).

The Nature estimator and the discrete variant (2024)

The 2024 Nature paper operationalizes equivalence through bidirectional entailment: answers share a cluster if each entails the other (pmc.ncbi.nlm.nih.gov). The probability-weighted estimator sums sequence probabilities within each cluster (pmc.ncbi.nlm.nih.gov). The discrete variant instead estimates each cluster probability as the proportion of sampled answers that belong to it (pmc.ncbi.nlm.nih.gov). That distinction matters for any audit, because the two estimators answer different questions. The discrete form uses only counts and therefore works where sequence probabilities are unavailable (pmc.ncbi.nlm.nih.gov). The probability-weighted form needs access to token probabilities.

The experimental protocol is explicit. The authors sample at temperature 1 with nucleus sampling (pmc.ncbi.nlm.nih.gov) and use ten generations to compute entropy (pmc.ncbi.nlm.nih.gov). Sentence-length experiments use 400 training and 400 test examples per dataset (pmc.ncbi.nlm.nih.gov). Semantic entropy averages an AUROC of 0.790 across 30 task and model combinations, ahead of naive entropy at 0.691 (pmc.ncbi.nlm.nih.gov), with the embedding regression baseline at 0.687 (pmc.ncbi.nlm.nih.gov). AUROC is the probability that a randomly chosen positive is ranked above a randomly chosen negative (people.inf.elte.hu), so 0.790 is a ranking property of the score and not an accuracy figure.

For long-form text the paper decomposes a passage into factual claims and generates new answers to each question (pmc.ncbi.nlm.nih.gov). Its biography set, FactualBio, is a collection of biographies of notable individuals (pmc.ncbi.nlm.nih.gov). In that setting discrete SE had better rejection accuracy than an adapted P(True) baseline until 20% of questions had been rejected (pmc.ncbi.nlm.nih.gov). The scope statements are the part most relevant here. The method is aimed at cases where LLMs are not consistently wrong because of erroneous training data (pmc.ncbi.nlm.nih.gov), and the Discussion says it does not address outputs that cause systematic reasoning errors or are systematically misleading the user (pmc.ncbi.nlm.nih.gov).

Probes, kernels and finer clustering (2024 to 2025)

Several follow-ups change what must be available, not what is measured. **Semantic entropy probes** approximate SE from the hidden states of a single generation (icml.cc), which removes multi-sample generation at test time. They need hidden-state access and a training set of SE-derived labels, which is the price of that saving. **Kernel language entropy** replaces hard clusters with graded semantic kernels and quantifies uncertainty with the von Neumann entropy (proceedings.neurips.cc), and is proved to generalize SE (proceedings.neurips.cc). **Semantic density** moves from a per-prompt score to a per-response confidence (proceedings.neurips.cc), motivated by SE being prompt-wise but not response-wise (proceedings.neurips.cc).

Efficiency-oriented variants lower the sampling cost; on this report's own reading, cost savings do not change what a single-cluster score means. The sources reviewed here do not claim to resolve the agreement-without-support case.

Finite samples and unobserved meanings (EACL, March 2026)

Evidential Semantic Entropy (EVSE) appeared in the proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (aclanthology.org). It starts from a finite-sample point: SE renormalizes cluster probabilities over what was observed, and the authors argue that this step discards information about unobserved answers (aclanthology.org). Using evidence theory, EVSE represents total ignorance arising from unobserved answers (aclanthology.org). Its requirements are explicit: it focuses on white-box methods using token-level probabilities (aclanthology.org), and the authors state it could not be adapted to a black-box setting (aclanthology.org). For this report the relevance is methodological. EVSE motivates a sample-budget sensitivity check. It does not establish that any answer is supported by a source.

Scoring without trusted reference evidence (September 2026 preprint)

The preprint titled "From Tokens to Semantics: Leveraging Complementary Signals for Hallucination Detection in Black-Box LLMs" (arxiv.org) studies the setting where no trusted context or reference document is available, using only signals accessible through black-box model APIs (arxiv.org). Three facts from it bear directly on routing decisions:

- **Supervision dependence.** Its stronger variants need labelled target-domain data, and zero-shot transfer is not evaluated (arxiv.org).
- **Dataset dependence.** The authors report that low false-positive-rate performance is strongly dataset-dependent (arxiv.org).
- **No universal winner.** Rankings are not statistically conclusive: the paper says its results do not establish statistically distinct superiority for an individual method on any single benchmark (arxiv.org).

This report treats the work as an arXiv preprint and does not establish its review or workshop status. The abstract page lists version 3 as the latest as of 2 November 2026 (arxiv.org).

Table 1: method ledger

Table 1 below summarizes what each method needs, what it measures and what it is documented not to cover. Rows and cells without a citation are this report's synthesis of the sources cited above.

Method	Access and samples	Labels needed	Measured target	Documented blind spot
Semantic entropy, probability-weighted (Kuhn 2023; Farquhar 2024)	Sampled answers plus sequence probabilities; ten generations in the Nature setup	None; unsupervised (iclr.cc)	Spread of meanings across clusters	Systematic errors and deception are outside scope

Method	Access and samples	Labels needed	Measured target	Documented blind spot
Discrete semantic entropy	Sampled answers only; cluster proportions	None	Same, from counts	Same; counts from a small sample can omit rare meanings (aclanthology.org)
Semantic entropy probes	Hidden states of one generation (icml.cc)	SE-derived binary labels on a training set	Approximation of SE	Inherits SE scope; needs white-box access
Kernel language entropy	Samples plus a semantic kernel (proceedings.neurips.cc)	None	Graded semantic spread	Requires multiple samples
Evidential semantic entropy	Token probabilities plus multiple samples	None	SE adjusted for unobserved meanings	Not black-box; not source support
No-reference black-box variants (2026 preprint)	Text, sometimes log-probabilities	Supervised variants need target-domain labels	Hallucination label from a judge protocol	Single-cluster cases; calibration varies by dataset (arxiv.org)
SelfCheckGPT	Sampled responses, black-box	None	Consistency of a response with samples	Consistency need not align with factuality

The table shows a pattern that repeats throughout this report. Every row measures something about the model's sampled behavior, and none takes the supplied source bytes as an input. Rows differ in whether they need probabilities, hidden states or labels, which determines whether a team can deploy them against a hosted model API at all.

Where Sampling Agreement Stops: Consistency, Correctness and Support

Semantic entropy versus self-consistency

The two terms are often conflated. **Self-consistency** in the sense of Wang et al. is a decoding strategy: it first samples a diverse set of reasoning paths instead of only the greedy one (iclr.cc) and selects the most consistent final answer. It reports accuracy gains such as +17.9% on GSM8K and +11.0% on SVAMP over greedy chain-of-thought decoding (iclr.cc). Its purpose is answer aggregation, not detection. **SelfCheckGPT** is the detection analogue: it samples extra responses and checks a response's consistency with them. It was evaluated on GPT-3 passages in which 761 of 1,908 annotated sentences (39.9%) were labelled major-inaccurate (aclanthology.org), so its headline numbers come from a single-domain biography set.

Semantic entropy sits between them. Like SelfCheckGPT it needs several samples and no source. Unlike token-level agreement, it clusters by meaning, so paraphrases do not inflate the score. The practical differences are in cost and scoring. On an independent benchmark of uncertainty methods across eleven tasks (direct.mit.edu), the best choice depended on the task, which is a reason to measure a detector on the target task and not to import a ranking.

On a retrieval-augmented benchmark, RAGTruth, a SelfCheckGPT baseline that samples three extra responses as references (arxiv.org) is one of several detector baselines, which shows that sampling-consistency was already evaluated against RAG labels. What is missing is a like-for-like comparison of clustered semantic entropy against source-support checks on identical labels.

What agreement can and cannot show

Consistency-based uncertainty has a documented limit. A survey of uncertainty quantification for LLMs states that consistency-based approaches face fundamental limitations, since consistency is not necessarily aligned with factuality (arxiv.org), while also noting that consistency is often a good predictor of factuality given enough samples. Both statements can hold: agreement is evidence about the model's distribution, and the distribution can be concentrated on a wrong or unsupported answer. Verbalized confidence is no escape. Xiong et al. observe that stated confidence appears to imitate human expressions (iclr.cc).

Source-grounded answers: support is its own label

Three lines of evidence show that source support has to be labelled directly.

First, the attribution literature separates support from truth by design. Attribution is framed as verifying generated text against an independent, provided source (direct.mit.edu). The AIS framework deliberately avoids factuality judgments, treating source quality as a complementary evaluation, and FActScore likewise defines truthfulness relative to a particular knowledge source.

Second, retrieval does not by itself settle support. RAGTruth annotates unsupported or contradictory claims relative to retrieved contents (aclanthology.org), which is a support-style label on retrieval-augmented responses and not a check against world knowledge.

Third, uncertainty estimators were examined under retrieval directly. Soudani et al. report that current uncertainty estimation methods cannot reliably estimate the correctness of LLM responses in the RAG setting (aclanthology.org), and that they systematically generated low uncertainty regardless of how relevant the context was to the query (aclanthology.org). This is the closest published result to the blind spot of interest, and it concerns correctness against ground truth, not support.

Checking support with entailment and judge methods

Support checkers have their own measurement properties, which the audit must record rather than assume. ALCE computes citation recall as 1 only if the concatenation of cited passages fully supports the segment (arxiv.org). Human agreement with its automatic checker was substantial for citation recall (Cohen's kappa 0.698) and moderate for precision (0.525) (aclanthology.org), so even a published support metric carries measurement error.

Metric choice changes conclusions. A 2025 re-evaluation reports that established detection methods show performance drops of up to 45.9% when the answer-matching metric is replaced with human-aligned judgment (aclanthology.org). That finding concerns labels, not the specific detectors named in this report, but it applies to any audit: the labelling protocol is part of the result.

Data Analysis and Evidence

What published numbers can and cannot support

Published evidence about semantic entropy is almost entirely no-reference. The Nature averages discussed above were reported without standard errors (www.nature.com). The preprint reports that the AUROC of one semantic-entropy variant rises from 0.468 to 0.750 on a single dataset as the generating model changes (arxiv.org), so a detector's score depends heavily on the generator and a threshold cannot be assumed to travel across models. The same paper says that its results do not establish statistically distinct superiority for an individual method on any single benchmark, and that low false-positive-rate performance is strongly dataset-dependent.

Two cautions follow for routing decisions. First, AUROC integrates over all thresholds, so a classifier with a higher AUROC can still perform worse in the region where a review budget places the threshold (people.inf.elte.hu). Second, when positives are rare, ROC plots can mislead and precision-recall views better predict future classification performance (pmc.ncbi.nlm.nih.gov).

The preprint does not claim transfer to RAG, and none is assumed here. Its detectors score without trusted reference evidence, and the abstract frames the setting as having no trusted context.

Synthetic arithmetic: why agreement alone is not identifiable

The following computations are synthetic. They are not observed benchmark results. They illustrate the empirical cluster entropy that SE computes from counts, using natural logarithms.

- **All ten samples in one cluster.** The cluster proportion is 1, so entropy is 0 nats. This holds identically when the cited source supports the answer and when the source contradicts it. The score is unchanged under the swap, so it cannot discriminate the two.
- **Nine of ten in one cluster, one in another.** Entropy is about 0.325 nats.
- **Five and five.** Entropy is ln 2, about 0.693 nats, the maximum for two clusters.
- **A rare meaning that is never drawn.** If a meaning has true probability 0.05 and draws were independent, the chance that ten samples all miss it is 0.95 to the tenth power, about 0.599; with five samples it is about 0.774. Real sampling may violate independence, so the figure is a lower-fidelity illustration, not an estimate. Plug-in entropy estimators are negatively biased, a general property of entropy estimation (www.stat.columbia.edu).

The point is narrow. Agreement among samples is evidence about the sampling distribution. It becomes evidence about support only if an independent check ties the sampling distribution to the source, which is what the audit's labels provide.

Table 2: the source-condition by cluster-count design

Table 2 below specifies the proposed paired audit design. It reports no results. Each row is a source condition; columns record what each cluster outcome would let an analyst conclude and which labels are adjudicated separately.

Source condition	Provenance of the case	Single-cluster outcome: what is recorded	Multi-cluster outcome: what is recorded	Labels adjudicated separately
(a) Sufficient support	Real source version, bytes and observation time preserved	Support expected; record any unsupported answer	Record whether extra clusters are paraphrase, refusal or distinct claims	Truth, support, citation resolves
(b) Decisive span removed	Authored edit; label **synthetic**	Key test: answer unsupported yet low entropy	Answer diversity without support	Truth, support, citation resolves
(c) Explicit counterevidence for the same entity and time	Authored edit; label **synthetic**	Key test: confident answer contradicting the supplied source	Mixed clusters, some contradicting	Truth, support, citation resolves
(d) Irrelevant context	Real or authored distractor	Answer from parametric memory; support absent	Spread reflects retrieval failure	Truth, support, citation resolves
(e) Legitimate multiple answers	Real question with several valid answers	Entropy zero but answer may omit valid options	Diversity is not automatically error	Truth, support, completeness

Authored edits in rows (b) and (c) are constructed to isolate the mechanism, so they must never be reported as real-world evidence. The interpretive point is the pairing: the same question and sampled meanings appear under several source conditions, so any change in cluster outcome or in support is attributable to the condition, and the low-entropy unsupported slice can be inspected case by case. This design is a proposal. It has not been executed, and the cell contents state what is recorded, not what is expected to be found.

Implementation Considerations and Process Changes

Build cases and preserve every sample

Select locally relevant questions that have independent reference judgments, then preserve the following for each case before any score is computed:

- **The exact question** and the prompt template that wrapped it.
- **Source bytes and version identifier**, so the snapshot the model saw can be reproduced.
- **Observation time** of the source, since support is relative to a version at a time.

- **Claim locators**, meaning the passage or span against which each claim is judged.
- **The adjudication policy** in force when labels were assigned.
- **Every candidate sample**, including refusals and failed generations, so that the analysis does not condition on convenient samples.

Keep the generator, prompt, decoding settings and clustering method fixed within a comparison. The published setups differ: temperature 1 with nucleus sampling in the Nature study, multiple samples in EVSE, ten by default in the preprint (arxiv.org). These are experimental settings and not requirements, so prespecify a series of sample counts, rerun clustering on nested samples, and preserve shared samples so comparisons are paired.

Audit the clustering, not just the score

The bidirectional-entailment judgment is where qualifiers, times and entities can be lost, and paraphrases that change the claim must not be merged. The audit therefore records:

- **The entailment model or prompt**, and whether the question and the source are included in the entailment check.
- **A blinded sample of cluster assignments**, each merge compared against the claim's qualifiers, time and entity.
- **Refusal as its own cluster**, separate from any substantive answer.
- **Per-answer evaluation**: each emitted answer is labelled, not only the majority sample, because the source-support label attaches to the answer a user actually sees.
- **The clustering variants tried**, so that sensitivity to the clustering policy is reportable.

Label three things independently, and keep disagreement

Reviewers work blind to the entropy score and label three things independently:

- **Source support**: whether the answer's claims follow from the identified source version.
- **Citation resolution**: whether each cited locator resolves to text containing the claim.
- **Externally established correctness**: whether the answer is true by an independent reference.

Retain disagreement as data. Agreement should be reported with a chance-corrected statistic: Cohen's kappa was designed because percent agreement cannot account for chance, and one methodological review proposes that kappa below 0.60 indicates inadequate agreement (pmc.ncbi.nlm.nih.gov). For more than two reviewers or missing ratings, Krippendorff's alpha accepts any number of observers with or without missing data (www.asc.upenn.edu). The ALCE human-agreement values reported earlier show the range that automatic and human support judgments can occupy.

Statistics: group, freeze, and report denominators

The statistical design follows from the grouped structure of the data:

- **Grouped splits.** Split development, calibration and test sets by underlying fact or source family so near-duplicates do not leak. scikit-learn's GroupKFold ensures each group appears exactly once in the test set across folds (scikit-learn.org), because dependent samples otherwise appear in both training and validation folds (scikit-learn.org).
- **Group-level intervals.** Compute intervals by question and source group, not by sampled answer. Default standard errors can greatly overstate estimator precision when errors are correlated within clusters (cameron.econ.ucdavis.edu), and with few clusters even a pairs cluster bootstrap did not eliminate over-rejection in the cited simulations (cameron.econ.ucdavis.edu).
- **Small-count intervals.** For small counts, the Wilson or Jeffreys intervals are recommended over the textbook Wald interval (repository.upenn.edu).
- **Declared denominators.** State the denominator of every rate.

Select the routing threshold on calibration data only, then freeze it for the held-out evaluation. If a bound on the error among accepted answers is required, the selective-classification formalism shows that risk can be traded off for coverage (papers.nips.cc). Record whether log-probabilities were available, because that decides which estimators are even feasible. Calibration of raw scores is a separate step: modern neural networks are often poorly calibrated (proceedings.mlr.press).

Table 3: review-budget operating points

Table 3 below defines the five quantities to report at each threshold, shown with a **(Hypothetical Example)** of 100 questions in which 20 answers are unsupported and 80 are supported. All figures are synthetic arithmetic, not results, and the baseline row assumes unsupported answers are spread uniformly through the review order. The quantities are:

- **Reviewed fraction**: questions sent to review over all questions.
- **Residual risk**: unsupported answers among those accepted, over all accepted.
- **Miss rate**: unsupported answers accepted, over all unsupported.
- **False-alarm rate**: supported answers sent to review, over all supported.
- **Baseline comparison**: plain source-support review at the same review budget.

Operating point	Reviewed / all questions	Unsupported accepted / accepted	Unsupported missed / all unsupported	False alarms / all supported
Threshold A (hypothetical)	30 / 100	8 / 70 (11.4%)	8 / 20 (40%)	18 / 80 (22.5%)
Threshold B (hypothetical)	50 / 100	4 / 50 (8%)	4 / 20 (20%)	34 / 80 (42.5%)
Plain support review, same budget as A (hypothetical, unordered)	30 / 100	14 / 70 (20%)	14 / 20 (70%)	24 / 80 (30%)
Skip review of low-entropy answers (policy under test)	By threshold	The quantity that decides the policy	Reported with its interval	Reported with its interval

The table separates two ideas that practitioners often merge. Missed unsupported answers over all unsupported answers is a sensitivity-style property of the detector. Unsupported accepted over accepted answers is the residual risk a downstream user inherits. A detector can look good on the first and still leave an unacceptable value on the second if the base rate of unsupported answers is high. The hypothetical Wilson 95% interval for 8 of 70 spans roughly 6% to 21% even before accounting for clustering of questions within source families, which illustrates how wide intervals are at realistic review-set sizes. The "plain review" baseline is essential: a detector earns its place only by beating independent source-support review at the same budget, and a published ordering of detectors by AUROC cannot show that.

Sample-count and clustering sensitivity

Report the same table at each nested sample count and for each clustering policy, including how abstentions and failed generations were handled. The expectation, which the audit tests and does not assume, is that more samples reveal more meanings and raise entropy for some answers, as EVSE's unobserved-mass argument suggests, but larger counts do not make samples statistically independent and do not prove coverage of all meanings. A plateau in the curve is a finding about that dataset and sampler, not a guarantee.

Implications and Future Directions

- **For routing decisions.** The defensible use of a sampling-based detector is triage under a declared budget. The audit's output is an operating point with its false-alarm rate on held-out supported answers and its residual unsupported rate among accepted answers. A threshold chosen on the Nature or preprint datasets does not transfer by default: the preprint stresses dataset-dependent calibration, and its supervised variants need target-domain labels. A zero-event result among accepted answers needs care too. With few accepted cases the interval remains wide: when no events are observed in N trials, the upper 95% limit is approximately 3/N (jamanetwork.com), so zero unsupported answers among 60 accepted answers would still be compatible with a rate near 5%. "None found" is not "none exist".
- **For what can be defended afterward.** What survives the audit is the observed support label, with its source identity and observation time, and the citation chain from the answer to a literal passage. An uncertainty score does not certify an individual answer. It can justify spending review effort on one answer rather than another.
- **Open questions.** No published study found for this report compares clustered SE with source-support checks on identical RAG labels and claim-level support annotations. The closest evidence is the RAG uncertainty study and the RAGTruth baselines, each with different datasets and labels. Whether SE adds routing value beyond a support checker is therefore open. Whether support-conditioned scoring (for example, sampling with the source present and absent) separates parametric answers from source-derived answers is also untested here.
- **Limitations of this report.** The audit design is a proposal and has not been run. Sample sizes, thresholds and clustering policies are placeholders for the study to set. The preprint is a preprint. EVSE is validated on short sequences according to its authors' stated scope, and it needs token probabilities. Several cited results use different judges and labels, so numbers across papers are not comparable.

Frequently Asked Questions (FAQs)

- **Does semantic entropy detect hallucinations?** It detects a subset: confabulations, defined as arbitrary and incorrect generations. It is not designed for consistently wrong outputs.
- **What are the main limitations of semantic entropy?** Systematic errors are out of scope; the score is zero when all samples share a cluster; clustering fidelity affects it; samples add cost; and unobserved meanings can be missed in finite samples.
- **How does semantic entropy differ from self-consistency?** Self-consistency selects the most consistent answer for accuracy. SE scores the spread of meanings across samples, after clustering by mutual entailment. SelfCheckGPT uses sample consistency for detection.
- **Can semantic entropy flag unsupported answers in RAG?** There is no direct evidence that it reliably does. Standard uncertainty estimators did not reliably estimate correctness in RAG, and support is a separate label from truth. Test it with the paired audit above.
- **How is uncertainty quantification used for hallucination detection more broadly?** Survey taxonomies group methods by access and signal; evaluations find task-dependent winners, and consistency is not necessarily aligned with factuality.
- **What is evidence-conditioned hallucination detection?** Scoring or labelling an answer relative to the evidence supplied to the model, for example entailment-based support or faithfulness conditioning, instead of only the model's sampling behavior.
- **How can LLM answer grounding be verified?** Decompose answers into atomic claims and judge each against the identified source, check that cited passages entail the claim, and confirm that the cited locator resolves.
- **How can unsupported claims in LLM answers be detected?** Use claim-level support judgments against the source, for example with faithfulness or NLI-based consistency checks, validated against human labels (this report's suggestion, not a result from a cited study).
- **What are semantic entropy probes?** Linear probes on hidden states trained to approximate SE from a single generation; they need white-box access and training data.

Conclusion

Semantic entropy measures the dispersion of sampled meanings. The original authors scope it to confabulations and say it does not guarantee factuality. The 2026 literature adds finite-sample refinements and documents single-cluster blind spots in no-reference settings, while the closest RAG study finds that standard uncertainty estimators are unreliable for response correctness. For a source-grounded answer, three labels matter and must be adjudicated separately: truth, support by the identified source version, and citation resolution. A detector can be useful for review routing without certifying any individual answer. The decision to add one, to calibrate a threshold, or to keep independent source-support review for low-entropy answers should rest on a paired audit with declared denominators, group-level intervals, a frozen threshold and a same-budget support-review baseline. Until that audit exists for a given source bundle, the supportable position is to retain independent review for low-entropy answers and to treat any entropy threshold as a hypothesis about review efficiency, not a statement about truth.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.nature.com/articles/s41586-024-07421-0
2. https://arxiv.org/abs/2609.02679
3. https://arxiv.org/html/2609.02679v3
4. https://aclanthology.org/2025.findings-acl.852.pdf
5. https://arxiv.org/pdf/2112.12870
6. https://aclanthology.org/2023.emnlp-main.741.pdf
7. https://aclanthology.org/2026.eacl-long.334
8. https://direct.mit.edu/coli/article/49/4/777/118691/Measuring-Attribution-in-Natural-Language
9. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
10. https://iclr.cc/virtual/2023/poster/11013
11. https://pmc.ncbi.nlm.nih.gov/articles/PMC11186750
12. https://people.inf.elte.hu/kiss/11dwhdm/roc.pdf
13. https://icml.cc/virtual/2024/39442
14. https://proceedings.neurips.cc/paper_files/paper/2024/hash/10c456d2160517581a234dfde15a7505-Abstract-Conference.html
15. https://proceedings.neurips.cc/paper_files/paper/2024/hash/f26d4fbaf7dfa115f1d4b3f104e26bce-Abstract-Conference.html
16. https://aclanthology.org/2026.eacl-long.334.pdf
17. https://arxiv.org/abs/2609.02679v3
18. https://iclr.cc/virtual/2023/poster/11718
19. https://aclanthology.org/2023.emnlp-main.557.pdf
20. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00737/128713/Benchmarking-Uncertainty-Quantification-Methods
21. https://arxiv.org/pdf/2401.00396
22. https://arxiv.org/pdf/2412.05563
23. https://iclr.cc/virtual/2024/poster/18135
24. https://aclanthology.org/2024.acl-long.585.pdf
25. https://arxiv.org/pdf/2305.14627
26. https://aclanthology.org/2023.emnlp-main.398.pdf
27. https://aclanthology.org/2025.emnlp-main.1761.pdf
28. https://pmc.ncbi.nlm.nih.gov/articles/PMC4349800
29. https://www.stat.columbia.edu/~liam/research/pubs/info_est-nc.pdf
30. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052
31. https://www.asc.upenn.edu/sites/default/files/2021-03/Answering%20the%20Call%20for%20a%20Standard%20Reliability%20Measure%20for%20Coding%20Data.pdf
32. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html
33. https://scikit-learn.org/stable/modules/cross_validation.html

34. https://cameron.econ.ucdavis.edu/research/Cameron_Miller_JHR_2015_February.pdf
35. https://repository.upenn.edu/handle/20.500.14332/47517
36. https://papers.nips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf
37. https://proceedings.mlr.press/v70/guo17a.html
38. https://jamanetwork.com/journals/jama/article-abstract/389714

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.