arXiv:2609.02679v2 [cs.CL] 4 Sep 2026

From Tokens to Semantics: Leveraging Complementary Signals for Black-Box LLM's Hallucination Detection

Urja Pawar
BNY
urja.pawar@bny.com

Rajitha Ramanayake
BNY
rajitha.ramanayake@bny.com

Nabeel Kemal
BNY
nabeel.kemal@bny.com

Owen O'Neill
BNY
owen.oneill@bny.com

Abhishek Mandal
BNY
abhishek.mandal@bny.com

Houssem Chatbri
BNY
houssem.chatbri@bny.com

Christopher Martin
BNY
Christopher.Martin@bny.com

Abstract

When LLMs support public-facing or high-stakes workflows, missed fabrications can harm users and institutions, while false alarms consume limited human-review capacity. When a hallucination-detection method cannot use a trusted reference, we study two signals accessible through black-box model APIs: semantic entropy, which measures disagreement among sampled response meanings, and uncertainty derived from token log-probabilities. Their failure modes can be complementary: semantic entropy becomes uninformative when responses form one semantic cluster, while token uncertainty can miss consistently confident errors. We extend token-based uncertainty detection by aggregating token-level signals across sampled responses through our **TopK** method, evaluate the hybrid CoCoA method, which combines target-response uncertainty with semantic dissimilarity, and propose and study two supervised methods: **Gated**, which routes single-cluster cases to an aggregated-token-feature classifier, and **Stacked**, which learns jointly from semantic uncertainty and broader token features. We evaluate seven benchmarks—including five public benchmarks (four text datasets and multimodal handwritten-cheque extraction) and two author-constructed benchmarks, Financial Summaries and Long-Text QA—using four language models. In our evaluation across models and datasets, Stacked gave the best performance in nearly half of the cases, while TopK and CoCoA remain competitive without supervised training labels, although their thresholds require careful calibration. No method is universally strongest. We therefore evaluate performance at false-positive-rate budgets from 1% to 15%, assess their sensitivity to generation and calibration choices, and examine variation across dataset characteristics.

1 Introduction

Large language models increasingly support information workflows in regulated organizations and public-facing services [Sajadieh et al., 2026]. In these settings, fabricated claims or incorrect figures can mislead downstream reviewers and users, while false alarms consume limited human-review

Preprint.

capacity. Reference-based methods can check outputs against a document or a context, but open-ended tasks may have no trusted context available at inference time. Proprietary models are commonly accessed through black-box APIs that expose generated text and sometimes token log-probabilities, but not hidden states [Kossen et al., 2025, Wang et al., 2025]. We therefore study whether these observable signals can identify hallucinated responses without using trusted reference evidence for scoring or requiring model-internal access.

Two API-visible signal families are especially relevant. Semantic entropy (SE) samples N responses, groups them by meaning, and measures uncertainty from their distribution across semantic clusters [Kuhn et al., 2023, Farquhar et al., 2024]. Token-log-probability methods instead estimate confidence within a response using sequence likelihood, perplexity, top-k entropy, and candidate margins (the differences between the highest and second-highest token log-probabilities) [Kadavath et al., 2022, Duan et al., 2024, Shapiro et al., 2026]. Because SE already requires multiple responses, we compute token signals for each response and aggregate them across the same samples. Our scalar **TopK** score combines mean top-k entropy with cross-response confidence variation to give an estimate of a model's uncertainty for a given question. We also evaluate CoCoA [Vashurin et al., 2025], a hybrid score that combines the confidence of the generated response with semantic dissimilarity to sampled alternatives.

The two signal families fail differently. SE becomes zero when all samples form one semantic cluster, even if they are incorrect. Token features can be informative in some such cases but can miss consistently confident hallucinations; conversely, semantic disagreement can expose errors that token probabilities do not. This complementarity motivates two supervised combinations. The *Gated cascade* uses semantic uncertainty when multiple clusters are present and routes single-cluster cases to an aggregated-token classifier. The *Stacked classifier* instead learns jointly from both signal families for every query (see Figure 1).

We evaluate seven benchmarks—including five public benchmarks (four text datasets and multimodal handwritten-cheque extraction) and two author-constructed benchmarks, Financial Summaries and Long-Text QA—using four language models where supported. In our evaluation across models and datasets, Stacked gave the best performance in nearly half of the cases, while TopK and CoCoA remain competitive without supervised training labels, although their thresholds require careful calibration. No method is universally strongest. We therefore evaluate performance at false-positive-rate budgets from 1% to 15%, assess sensitivity to generation and calibration choices, and examine variation across dataset characteristics (see Appendix E). We document the construction of our two benchmarks for reproducibility but do not present them as openly released datasets.

Research questions. We ask where semantic and token-log-probability signals fail across datasets; whether their complementary information can improve black-box hallucination detection; and how consistently TopK, Gated, and Stacked perform across datasets, models, false-positive-rate budgets, and data characteristics. Our findings provide conditional guidance for the evaluated setting types.

2 Complementary Black-Box Hallucination Signals

2.1 Related work

Semantic Entropy. Semantic entropy (SE) [Kuhn et al., 2023, Farquhar et al., 2024] draws N responses to a prompt x, groups them into semantic equivalence classes $\mathcal{C} = \{C_1, \ldots, C_K\}$, and measures uncertainty over the empirical class distribution:

$$\mathrm{SE}(x) = -\sum_{k=1}^{K} \hat{p}(C_k) \log \hat{p}(C_k), \qquad \hat{p}(C_k) = |C_k|/N. \tag{1}$$

Standard SE has two limitations: finite sampling may omit plausible semantic classes, while hard clustering discards graded similarities. Alphabet-size corrections estimate classes unobserved in the sample [McCabe et al., 2026]. Pairwise-similarity methods retain graded information: Kernel Language Entropy uses a semantic-similarity kernel, while Semantic Nearest Neighbor Entropy (SNNE) uses nearest-neighbour similarities without discrete classes [Nikitin et al., 2024, Nguyen et al., 2025]. Spectral Uncertainty, which we evaluate, uses the response-similarity graph spectrum to retain graded information and separate aleatoric and epistemic uncertainty [Walha et al., 2025].

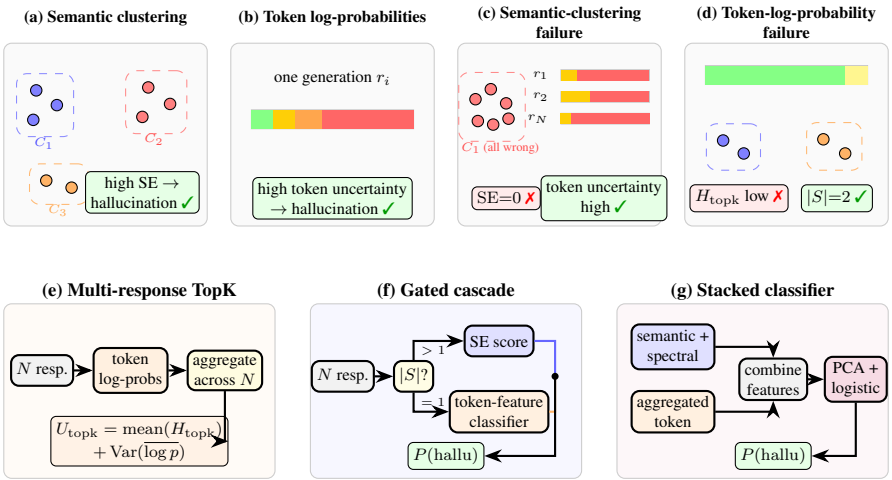

Figure 1: **Complementary black-box signals and methods. (a–d)** Semantic and token uncertainty expose complementary failures. **(e–g)** TopK aggregates token uncertainty across responses, Gated routes by cluster count, and Stacked learns jointly from semantic and token features.

Semantic Energy augments semantic classes with penultimate-layer logits but requires white-box access [Ma et al., 2025]. Semantically Diverse Language Generation (SDLG) elicits diverse alternatives from one model, while cross-model disagreement compares similarities within and across models [Aichberger et al., 2025, Hamidieh et al., 2026]. These methods enrich semantic variation but may provide limited evidence when the same wrong meaning repeats, motivating the token-level evidence considered next.

Token-level uncertainty. Token-probability methods assess confidence within a generation rather than disagreement among response meanings. Basic approaches use sequence likelihood or predictive entropy, while self-evaluation uses the probability assigned to TRUE as confidence [Malinin and Gales, 2021, Kadavath et al., 2022]. Shifting Attention to Relevance (SAR) weights uncertainty associated with semantically important tokens and responses [Duan et al., 2024]. EAS accumulates entropy along a reasoning sequence, whereas Hallucination Assessment via Log-probs as Time Series (HALT) learns temporal patterns from top-k token log-probabilities [Zhu et al., 2025, Shapiro et al., 2026]. Unlike HALT, which models uncertainty across token positions within one response, our methods aggregate token evidence across the multiple responses already sampled for semantic entropy; consistently confident hallucinations nevertheless remain difficult to identify.

Hybrid and higher-access methods. CoCoA is a closely related hybrid method. It multiplies a target response's scalar model uncertainty by its average semantic dissimilarity to sampled alternatives, deriving the combination through a Minimum Bayes Risk formulation [Vashurin et al., 2025]. We instead aggregate broader token-probability features across all N responses and combine them with query-level semantic signals.

2.2 Proposed methods

The different failure modes of semantic and token signals motivate three methods using the same N sampled responses and API-exposed log-probabilities. TopK aggregates token uncertainty across responses. Gated and Stacked combine token and semantic evidence: Gated routes queries based on whether multiple semantic clusters are present, whereas Stacked learns jointly from both signal families. Note that the Gated and Stacked methods require labelled training data.

Multi-response token representation. For each response, we extract token confidence, sequence confidence, positional confidence changes, and response-length features, then summarise their distributions across the N responses using means, variances, quantiles, ranges, and extreme values. TopK uses only mean top-k entropy and across-response confidence spread, whereas Gated and Stacked use the broader aggregated feature vector. For analysing which types of features impact detection, we group log-probabilities, candidate margins, and entropy as **token confidence**; positional

confidence changes and response length as **response dynamics**; cross-sample spreads and extremes as **across-response aggregation**; and response diversity and Von Neumann entropy as **semantic diversity**.

Gated cascade. The Gated cascade uses the semantic-cluster count K for routing. When $K \geq 2$, it returns the selected semantic entropy score; when $K = 1$, empirical SE is zero, so it uses a logistic-regression classifier over the aggregated token features:

$$\text{score}(x) = \begin{cases} \text{SE}_m(x), & K \geq 2, \\ \sigma\big(\mathbf{w}^\top \mathbf{f}_{\text{tok}}(x) + b\big), & K = 1. \end{cases} \tag{2}$$

Here, $\mathbf{f}_{\text{tok}}(x)$ is the token-feature vector aggregated across the N responses, and σ converts the classifier output into a hallucination probability. As described in the following section (Table 1), *Gated (Hybrid)* uses Hybrid SE on the semantic branch, whereas *Gated (Spectral)* uses Von Neumann entropy. They share the same token classifier and differ only in their semantic entropy score.

Stacked classifier. The Stacked classifier combines the aggregated token profile with one semantic feature block—*Hybrid, Spectral,* or *Von Neumann*—that includes the cluster count. The features are standardised, reduced using principal component analysis (PCA), and passed to L_2-regularised logistic regression, allowing both signals to influence every prediction (for details on feature inputs, see Appendix C.4).

3 Methodology and Experimental Setup

For each query, we compute several semantic uncertainty scores and the multi-response token representation from Section 2.2, evaluating the measures independently as well as jointly via the Gated and Stacked approaches. This section defines the variants, datasets, models, training procedure, and metrics.

3.1 Methods and Variants

Table 1 provides an overview of the selected methods and how their measures are used through the Gated and Stacked approaches.

Table 1: **Overview of the 13 evaluated methods.** Descriptions summarise the uncertainty score or classifier inputs used by each method.

Method	Description
Standard SE	Empirical entropy of the semantic-cluster frequencies in Eq. 1 [Kuhn et al., 2023, Farquhar et al., 2024].
SE UEigV	Adjusts the finite-sample class estimate using eigenvalues of the response-similarity graph [McCabe et al., 2026].
SE Hybrid	Combines UEigV with a Good–Turing correction based on singleton classes, using the larger estimate [McCabe et al., 2026].
SE Von Neumann	Von Neumann entropy of a density matrix derived from graded response similarities [Walha et al., 2025].
TopK	Combines mean top-k entropy with confidence variation across responses; uses no supervised training labels but requires threshold calibration.
Spectral Epistemic	Epistemic component obtained by decomposing spectral uncertainty [Walha et al., 2025].
CoCoA SP	Multiplies target sequence-level uncertainty by its average semantic dissimilarity to sampled alternatives; uses no supervised training labels but requires threshold calibration [Vashurin et al., 2025].
CoCoA PPL	Uses the length-normalised target uncertainty within the same CoCoA combination; uses no supervised training labels but requires threshold calibration [Vashurin et al., 2025].
Gated Hybrid	Uses Hybrid SE for multiple-cluster queries and an aggregated-token classifier for one-cluster queries.
Gated Spectral	Uses Von Neumann entropy for multiple-cluster queries and the same aggregated-token classifier for one-cluster queries.
Stacked Hybrid	Classifier combining the token vector with Hybrid SE, estimated semantic alphabet size, and cluster count.
Stacked Spectral	Classifier combining the token vector with total spectral entropy, effective rank, epistemic uncertainty, and cluster count.
Stacked Von Neumann	Classifier combining the token vector with Von Neumann entropy and cluster count.

We embed the N responses using `text-embedding-3-large` and use pairwise cosine similarity to measure their semantic similarity. For Standard SE, UEigV, and Hybrid SE, a dataset-specific threshold τ determines whether a response is assigned to an existing semantic class. We select τ using separate validation data and fix it before five-fold cross-validation; the validation examples are excluded from the evaluation folds. Von Neumann and spectral scores use the same similarity matrix directly (full implementation details are provided in Appendix A.1).

Table 2: **Dataset profiles and hyperparameters.** n is the number of valid labelled examples used in five-fold cross-validation. Complexity values are qualitative 0–5 ratings defined below. τ is the cosine-similarity clustering threshold, and C is the L_2 regularisation strength for Gated (G) and Stacked (S).

			Complexity					C	
Dataset	**Description**	n	**Ctx**	**Ans**	**Reas**	**Dom**	τ	**G**	**S**
AA Omni Finance	Finance/law knowledge eval; 4-point grading [Team, 2025]	200	0	1	2	4	0.95	1.0	1.0
AmbigQA	Ambiguous open-domain QA; multiple valid answers [Min et al., 2020]	198	0	3	1	0	0.95	0.1	0.1
HotpotQA	Multi-hop reasoning over Wikipedia [Yang et al., 2018]	199	2	1	4	0	0.85	0.1	0.1
Cheque Generation	Handwritten cheque VQA; numeric extraction [Verma, 2024]	154	5	2	1	3	0.85	0.1	1.0
Financial Summaries	Synthetic finance texts with confusable entities (author-constructed)	171	1	4	3	3	0.90	1.0	0.1
Long-Text QA	Multi-part questions over regulatory filings (author-constructed)	30	3	5	3	4	0.90	10.0	1.0
SQuAD	Extractive reading comprehension [Rajpurkar et al., 2016]	200	2	1	1	0	0.80	0.1	0.1

Ctx (context): 0 = none, 1 = inline, 2 = short doc, 3 = long doc, 4 = multi-doc, 5 = multi-modal. **Ans** (answer): 0 = binary, 1 = short factual, 2 = structured, 3 = multi-answer, 4 = paragraph, 5 = multi-part. **Reas** (reasoning): 0 = lookup, 1 = single-hop, 2 = multi-hop, 3 = synthesis, 4 = judgement, 5 = creative. **Dom** (domain): 0 = general, 1 = semi-general, 2 = domain-specific, 3 = expert, 4 = niche expert, 5 = regulated.

3.2 Datasets

We evaluate seven benchmarks spanning factual, ambiguous, multi-hop, extractive, long-context, financial, and multimodal settings. Five are public, while Financial Summaries and Long-Text QA are author-constructed. Table 2 summarises the datasets, citations, task descriptions, and resulting sample sizes. It also characterises each dataset by context complexity, answer complexity, reasoning depth, and domain specificity. These human-assigned ratings use a 0–5 ordinal scale and provide qualitative context (further details on dataset construction and licenses are in Appendix C).

3.3 Models, Training, and Evaluation

Models. We evaluate GPT-4.1-mini, GPT-5.1, GPT-5.4, and Llama 3.3 70B where supported. Of the 28 possible dataset–model pairs, 26 contain both label classes and yield estimates of the area under the receiver operating characteristic curve (AUROC). The labelled Llama 3.3 70B outputs for Long-Text QA contain only the hallucinated class, so AUROC is undefined, while Cheque Generation requires vision capability and is not evaluated with the Llama model.

Generation and training. The default setting uses $N = 10$ responses per prompt, temperature 1.0, and $k = 5$ requested log-probability candidates per token. GPT-5.4 returned only one candidate, giving an effective $k = 1$ for token-based methods. All transformations and classifiers are fitted on the training portion of each stratified five-fold split and evaluated on held-out predictions.

Evaluated methods. We compare thirteen methods: four SE baselines (Standard SE, SE UEigV, SE Hybrid, and SE Von Neumann); TopK, Spectral Epistemic, CoCoA SP, and CoCoA PPL, which do not use supervised training labels; and five supervised methods (Gated Hybrid, Gated Spectral, Stacked Hybrid, Stacked Spectral, and Stacked Von Neumann). CoCoA SP and CoCoA PPL combine sampled-response consistency with sequence-probability (SP) and perplexity confidence (PPL), respectively.

Evaluation metrics. For each dataset, we compute AUROC separately within each of the five held-out folds and report the mean across folds in the primary model tables. We separately pool the out-of-fold predictions from all five folds into a single set, which we use for the cross-model comparisons, query-level bootstrap confidence intervals, and other supporting analyses. We also report true-positive rate (TPR) at false-positive-rate (FPR) budgets from 1% to 15%. These operating points trade hallucination coverage against false positives, review capacity, and the relative costs of

Table 3: **GPT-5.4 AUROC by method and dataset.** Values are mean held-out AUROC over five folds. Text datasets use $N = 10$ and Cheque Generation uses $N = 20$. The API returned one candidate per token, giving an effective $k = 1$ for token-based methods. Bold marks the highest value per dataset.

Method	AA Fin.	AmbigQA	Summ. Fin.	HotpotQA	SQuAD	Long QA	Cheque
Standard SE	0.507	0.623	0.492	0.624	0.481	0.620	0.500
SE UEigV	0.506	0.630	0.492	0.623	0.481	0.595	0.500
SE Hybrid	0.507	0.634	0.492	0.628	0.481	0.595	0.500
SE Von Neumann	0.538	0.629	**0.750**	0.668	0.549	0.620	0.774
TopK	**0.721**	0.630	0.560	0.709	**0.745**	0.685	**0.812**
Spectral Epistemic	0.504	0.615	0.492	0.631	0.481	0.620	0.500
CoCoA SP	0.580	**0.670**	0.694	0.681	0.621	**0.725**	0.770
CoCoA PPL	0.586	0.648	0.647	0.702	0.608	0.620	0.770
Gated Hybrid	0.611	0.649	0.409	0.691	0.655	0.435	0.809
Gated Spectral	0.583	0.638	0.409	0.650	0.655	0.435	0.809
Stacked Hybrid	0.713	0.611	0.516	0.710	0.678	**0.725**	0.802
Stacked Spectral	0.714	0.629	0.630	**0.720**	0.660	0.625	0.808
Stacked Von Neumann	0.714	0.620	0.541	0.701	0.659	0.675	0.804

the two error types. The leading method's 95% confidence interval overlaps at least one competitor on every dataset, so close rankings are not conclusive (Appendix F.2 reports confidence intervals, and Appendix F.1 reports significance tests).

Sensitivity analyses. We vary the response count $N \in \{3, 5, 7, 10, 15, 20\}$, temperature in $\{0.3, 0.5, 0.7, 1.0, 1.2\}$, returned candidates $k \in \{1, \ldots, 5\}$, clustering threshold $\tau \in \{0.80, 0.85, 0.90, 0.95, 0.98\}$, and regularisation strength $C \in \{0.1, 1, 10\}$. Cross-model comparisons reuse the selected τ and C values. Section 4 summarises the findings (see Appendix E for the complete ablation results).

4 Results and Discussion

Across seven benchmarks and four models, we find complementary failure modes for semantic disagreement and token uncertainty. Among hallucinated queries, the percentage of queries where sampled responses form one semantic cluster ranges from 39% on AmbigQA to 99% on Financial Summaries, forcing standard SE to zero. Within these single-cluster cases, median TopK uncertainty is higher for hallucinated than non-hallucinated queries on six of seven datasets. However, depending on the dataset, 21–56% of hallucinated single-cluster queries have lower TopK uncertainty than the median non-hallucinated query, making TopK non-informative in those cases. Thus, token evidence remains informative in many cases where semantic disagreement disappears, although both signals can fail when a hallucination is semantically consistent and generated with high token confidence (see Appendix D.1 for the failure mode analysis).

4.1 Performance patterns across methods, datasets, and models

We group the methods into five families: semantic or spectral scores, TopK, CoCoA scores, Gated classifiers, and Stacked classifiers. Supervised classifiers lead four of seven GPT-4.1-mini comparisons, four of five valid Llama comparisons, and three of seven GPT-5.1 comparisons. Methods without supervised training labels—TopK, CoCoA, and semantic or spectral scores—lead or share the lead on six of seven GPT-5.4 comparisons. The GPT-5.4 results suggest that graded semantic measures can be quite informative but this pattern is not uniform and does not establish that model capability causes the improvement. For example, SE Von Neumann changes sharply with the generating model, rising from 0.468 to 0.750 on Financial Summaries and from 0.581 to 0.774 on Cheque Generation between GPT-4.1-mini and GPT-5.4. AmbigQA is the clearest exception to the supervised pattern, with CoCoA, which does not use supervised training labels, leading all four valid model comparisons. Among models with broad text-dataset coverage, Llama shows the weakest overall separation: the best method for each of its five valid comparisons averages about 0.64 AUROC. Together, these shifts show that the most informative uncertainty signal depends on the generating model as well as the dataset.

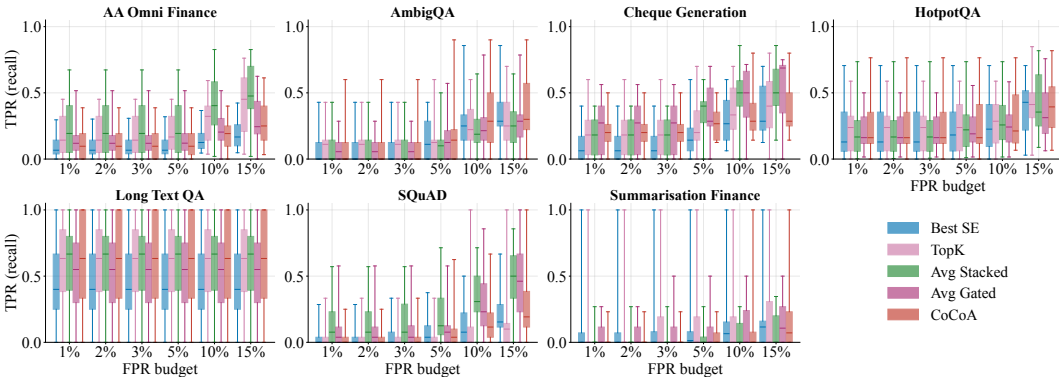

Figure 2: **TPR at fixed FPR budgets.** Boxes pool five-fold TPR values across available model runs at fixed calibration. Text runs use $N = 10$ and $T = 1.0$; Cheque Generation uses $N = 20$ and includes GPT-4.1-mini, GPT-5.1, and GPT-5.4. Long-Text QA has $n = 30$.

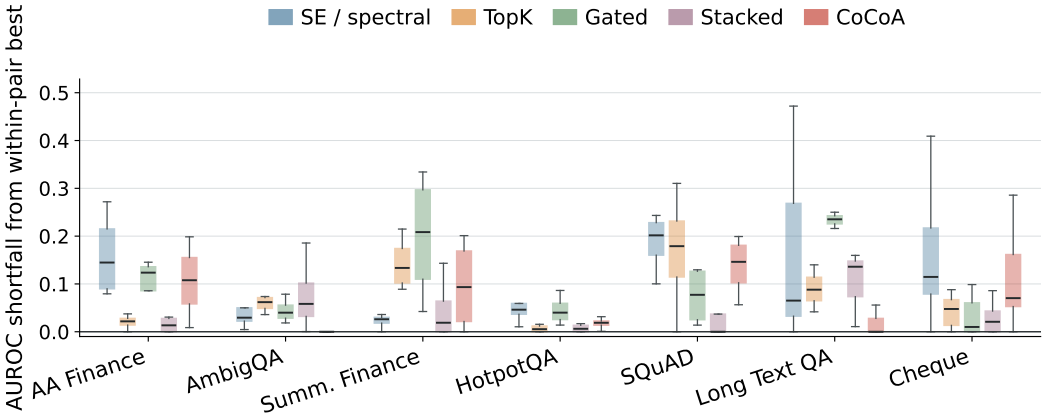

Figure 3: **Method-family AUROC shortfall.** Shortfall is measured from the best pooled out-of-fold AUROC in the same dataset–model setting; lower is better. The 26 settings comprise 23 text settings and three vision-capable Cheque Generation settings.

Figure 2 shows that low-FPR performance is strongly dataset-dependent. At strict 1–3% FPR budgets, the leading methods on the datasets showing clearer separation are predominantly token-based or supervised combinations: TopK leads on AmbigQA and HotpotQA, Stacked on AA Omni Finance and SQuAD, and Gated on Cheque Generation. At higher budgets, CoCoA becomes strongest on AmbigQA, while the leading family also changes on HotpotQA and Cheque Generation. Stacked remains strongest on AA Omni Finance and SQuAD. Increasing the FPR budget from 5% to 15% substantially improves hallucination coverage on AA Omni Finance, SQuAD, and Cheque Generation, but produces little improvement on Financial Summaries or Long-Text QA. Thus, both the preferred method family and the benefit of relaxing the threshold depend on the dataset.

Across the 26 comparisons, Stacked leads or shares the lead in 11, CoCoA in seven, TopK in five, Gated in three, and semantic or spectral scores in one; the counts sum to 27 because Stacked and CoCoA tie in one setting. Because the leading method changes across models, datasets, and FPR budgets, win counts alone do not show how consistently a method performs. We therefore also measure how far each family falls below the best method in the same model–dataset comparison. Figure 3 shows that Stacked remains consistently close to the leader: its strongest variant is within 0.05 AUROC of the best method in 20 of 26 comparisons and within 0.02 in 16. TopK and CoCoA have the next-smallest median shortfalls, indicating that they remain competitive in more specific settings even when they do not lead (see Appendix E.1 for the model-family ablation results).

4.2 Why performance varies across datasets

Figures 4 and 5 connect these performance differences to the observed uncertainty signals. Figure 4 reports the strongest absolute Cohen's d, the standardised mean difference, within each Stacked feature group. The grid contains up to 23 canonical and ablation settings per dataset and these are distinct from the 26 model–dataset comparisons above. Token-confidence or response-dynamics features provide the strongest separation in the majority of records. Because the available records are not independent, these frequencies are descriptive rather than evidence of cross-task generalisation.

AmbigQA provides the clearest semantic-disagreement pattern. For "How many jury members [are] in a criminal trial?", the accepted answers include 6, 7, 12, and 15 because the answer depends on jurisdiction. Variation among sampled responses is therefore meaningful rather than automatically erroneous. Ambiguity itself is not evidence of hallucination, since several interpretations may be valid. AA Omni Finance instead provides a token-confidence pattern centred on precise entities, dates, and amounts. For example, one question expects "December 15, 2022," while the hallucinated response changes only the year to 2021. Token confidence leads 15 of 23 settings. Figure 5 shows higher entropy for hallucinated responses across most token positions. Cheque Generation presents a more local pattern: a visually confusable digit can change the extracted answer without substantially changing its overall meaning. Although its positional profiles overlap, the difference between the two most probable token candidates can retain uncertainty at the decisive digit (see Appendix D.2 for the complete token-level analysis).

HotpotQA exhibits a mixed-evidence pattern. Its later-position entropy gap is consistent with uncertainty accumulating as the model combines supporting facts; an error can change token confidence, response meaning, or both. The remaining datasets exhibit weak or potentially confounded patterns. Financial Summaries combines similar fictional names and nearby figures, while substitutions remain semantically consistent and confidently generated; both signals therefore provide weak separation. In SQuAD, response-dynamics features lead 22 of 23 settings, with response length strongest. Its short extractive answers make longer responses and inverted early-position entropy informative, but these associations may reflect task format rather than general hallucination evidence. Long-Text QA shows a suggestive later-position gap because the final sub-question is deliberately unanswerable.

4.3 Conditional method and threshold selection

The results support selecting both the method and its operating point for the target setting. With labelled target-domain data, Stacked is a strong starting point because it uses semantic and token evidence for every query and usually remains close to the leading method; without supervised training labels, TopK and CoCoA are useful alternatives, but their thresholds still require calibration on representative data. The ablations show where additional API information is worth its cost. On AA Omni Finance, increasing the retained token candidates from one to three raises TopK's performance from 0.621 to 0.721 AUROC and average Stacked AUROC from 0.634 to 0.741, consistent with alternatives helping around precise dates, amounts, and entities. The same change leaves average Stacked essentially unchanged on SQuAD (0.767 to 0.768) and is non-monotonic on Long-Text QA. Sampling ten responses gives the highest mean AUROC across the six shared text datasets, but individual optima vary. The temperature sweep also shows opposite dataset-level preferences. For average Stacked, AmbigQA peaks at 0.700 AUROC with a temperature of 0.5 and falls to 0.461 at 1.2, whereas HotpotQA reaches its highest AUROC of 0.706 at 1.2, compared with 0.634 at 0.5. This is consistent with HotpotQA's single-cluster rate falling from 0.89 to 0.79 as temperature rises, giving semantic-diversity features more multi-cluster cases; AmbigQA does not follow the same pattern, since its single-cluster rate is highest, not lowest, at the temperature where its AUROC peaks. Thus, increasing sampling randomness does not consistently improve separation: the useful degree of response variation depends on the dataset. Overall calibration is stable for Stacked on AA Omni Finance and SQuAD but much more variable on the small Long-Text QA dataset. These results motivate choosing response count, token-candidate count, temperature, and threshold jointly with the dataset, model, available labels, false-positive cost, and review capacity (see Appendix E for the complete generation and calibration sweeps).

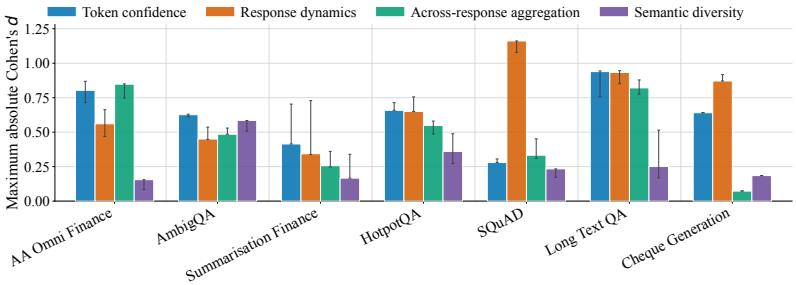

Figure 4: **Discriminative associations of feature groups by dataset.** Bars show the median, across ablation runs, of the largest absolute Cohen's d within each group; whiskers show the interquartile range.

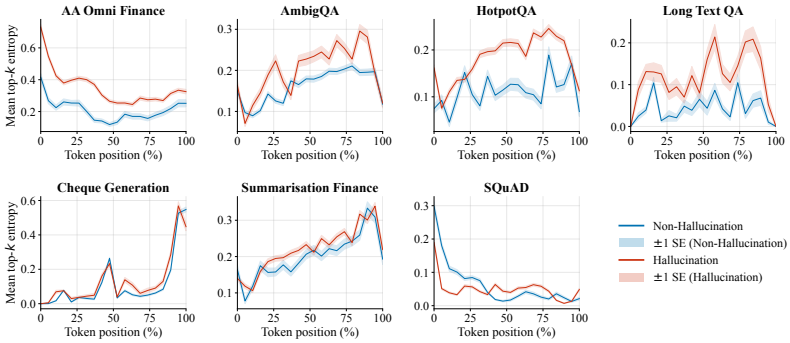

Figure 5: **Positional top-k entropy profiles.** Mean top-k entropy over normalised token position with ± 1 standard-error bands. Separation is persistent for AA Omni Finance, emerges later for several QA tasks, is weak for Cheque and Financial Summaries, and is inverted near the opening of SQuAD responses.

5 Limitations

Gated and Stacked require labelled target-domain hallucination data; zero-shot transfer is not evaluated. Long-Text QA contains only 30 questions. On every dataset, the leading method's confidence interval overlaps that of at least one competitor. Token-based methods require API-exposed log-probabilities, and GPT-5.4 returned only one candidate per token, limiting the available alternative-token information. In addition, TopK's confidence-spread component is heavy-tailed and produces extreme outliers, affecting both the scalar score and learned feature vectors. We therefore use median-based summaries in the corresponding appendix analysis, but robust alternatives require further study (Appendix F reports the significance tests and bootstrap confidence intervals).

6 Conclusion

We studied black-box hallucination detection using semantic entropy and token log-probabilities from sampled responses, without requiring reference evidence for scoring. Across seven benchmarks, four models, and 26 available model–dataset comparisons, the failure-mode analysis showed that the two signal families provide complementary information, although both can fail when hallucinations are semantically consistent and generated with high token confidence. Our fixed-FPR analysis makes explicit the trade-off among hallucination coverage, false positives, review capacity, and error costs. Stacked led or shared the lead in 11 of 26 comparisons and remained within 0.05 AUROC of the best method in 20 of 26, making it a strong starting point when labelled target-domain data are available. When labelled data are unavailable for fitting a supervised detector, TopK and CoCoA are competitive alternatives, although their deployment thresholds require calibration on representative data. Performance varied across datasets and generating models, and changes to response count, temperature, and token-candidate count produced no uniform improvement. These findings support selecting the method, generation configuration, and operating threshold jointly for the target setting.

Acknowledgments and Disclosure of Funding

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. Improving uncertainty estimation through semantically diverse language generation. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=HSi4VetQLj. arXiv:2406.04306.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.276. URL https://aclanthology.org/2024.acl-long.276/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=lOoRJo8xWy. arXiv:2604.17112.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=YQvvJjLWX0.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve. arXiv:2302.09664.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. Semantic energy: Detecting LLM hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*, 2025. URL https://arxiv.org/abs/2508.14496.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *The Ninth International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m. arXiv:2002.07650.

Lucas H. McCabe, Rimon Melamed, Thomas Hartvigsen, and H. Howie Huang. Estimating semantic alphabet size for LLM uncertainty quantification. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=uYK6GPVg1O. arXiv:2509.14478.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4530–4540. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.234. URL https://aclanthology.org/2025.findings-acl.234/. arXiv:2506.00245.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://openreview.net/forum?id=j2wCrWmgMX. arXiv:2405.20003.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Sha Sajadieh, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Lapo Santarlasci, Juan Pava, Nestor Maslej, Russ Altman, Erik Brynjolfsson, Carla Brodley, Jack Clark, Virginia Dignum, Vipin Kumar, James Landay, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Elham Tabassi, Russell Wald, Toby Walsh, and Dan Weld. The AI index 2026 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2026. URL https://hai.stanford.edu/ai-index/2026-ai-index-report.

Ahmad Shapiro, Karan Taneja, and Ashok Goel. HALT: Hallucination assessment via log-probs as time series. *arXiv preprint arXiv:2602.02888*, 2026. URL https://arxiv.org/abs/2602.02888.

Artificial Analysis Team. Artificial analysis omniscience, 2025.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Uncertainty quantification for LLMs through minimum Bayes risk: Bridging confidence and consistency. *arXiv preprint arXiv:2502.04964*, 2025. doi: 10.48550/arXiv.2502.04964. URL https://arxiv.org/abs/2502.04964.

Aniket Verma. Handwritten cheque VQA dataset. https://huggingface.co/datasets/aniketVerma07/handwritten_cheque_vqa_dataset, 2024. Apache-2.0 License.

Nassim Walha, Sebastian G. Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. Fine-grained uncertainty decomposition in large language models: A spectral approach. In *NeurIPS 2025 Workshop on Reliable ML from Unreliable Data*, 2025. URL https://openreview.net/forum?id=8IeWglLF7n. arXiv:2509.22272.

Tuo Wang, Adithya Kulkarni, Tyler Cody, Peter A. Beling, Yujun Yan, and Dawei Zhou. GENUINE: Graph enhanced multi-level uncertainty estimation for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20522–20541, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1119. URL https://aclanthology.org/2025.findings-emnlp.1119/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Yongfu Zhu, Lin Sun, Guangxiang Zhao, Weihong Lin, and Xiangzheng Zhang. Uncertainty under the curve: A sequence-level entropy area metric for reasoning LLM. *arXiv preprint arXiv:2508.20384*, 2025. URL https://arxiv.org/abs/2508.20384.

Appendix

The appendix is organised into six parts. Appendix A provides further details on the evaluated uncertainty estimation methods, including CoCoA and the Gated and Stacked algorithms. Appendix B describes the experimental implementation, including the embedding model, clustering configuration, and classifier settings. Appendix C documents the datasets, token-level features, dataset-construction procedures, and prompts used to obtain hallucination labels. Appendix D presents extended diagnostic results, beginning with the complementary failure modes of semantic and token uncertainty and followed by ROC and cross-model analyses. Appendix E reports the response-count, temperature, top-k, model, and calibration ablations, while Appendix F provides statistical significance tests and bootstrap confidence intervals. CoCoA is included in direct comparisons, ROC curves, significance tests, bootstrap intervals, and the top-k table, and appears as an invariant reference in calibration because it uses neither τ nor C. It is omitted from the response-count and temperature sweeps because the required main-answer embeddings are unavailable, and from the externally scored Cheque sweep because no CoCoA scores were produced.

A Uncertainty detection methods

This section provides the methodological details omitted from the main text for brevity. We first describe the semantic and spectral uncertainty measures, followed by TopK, CoCoA, and the Gated and Stacked methods.

A.1 Semantic and spectral uncertainty measures

All semantic baselines are computed from the same N sampled responses and their pairwise cosine-similarity matrix. Standard SE uses the empirical semantic-cluster distribution defined in Eq. 1. The remaining variants either correct for semantic classes that may be unobserved in a finite sample or retain graded similarities instead of reducing each response pair to a binary clustering decision.

The Good–Turing correction estimates the semantic alphabet size from the number of observed clusters K and singleton clusters n_1:

$$\widehat{K}_{\mathrm{GT}} = \frac{KN}{N - n_1}.$$

UEigV instead estimates the effective number of clusters by counting eigenvalues of the normalised similarity-graph Laplacian below a threshold ϵ. Hybrid uses the larger of the Good–Turing and UEigV estimates, falling back to UEigV when every observed cluster is a singleton, and recomputes entropy using the corrected alphabet size [McCabe et al., 2026]. Von Neumann entropy avoids a discrete alphabet-size estimate and computes uncertainty directly from the spectrum of the response-similarity matrix [Walha et al., 2025].

A.2 TopK

For response i, let $H_i^{(k)}$ denote its mean token entropy after renormalising the returned candidate probabilities over the available k candidates at each token position, and let $\bar{\ell}_i$ denote its mean chosen-token log-probability. The TopK score is

$$U_{\mathrm{TopK}}(x) = \frac{1}{N}\sum_{i=1}^{N} H_i^{(k)} + \mathrm{Var}_{i=1,\ldots,N}\left(\bar{\ell}_i\right).$$

Algorithm 1 Gated and Stacked uncertainty detection

Require: Query x, language model p_θ, response count N, semantic variant m, fitted Gated classifier $\mathcal{F}_G$, fitted Stacked pipeline $\mathcal{F}_S$, method d
Ensure: Uncertainty score s

Sample responses $\{y^{(i)}\}_{i=1}^{N} \sim p_\theta(\cdot \mid x)$ and retain their token log-probabilities
Embed the responses and compute their pairwise cosine similarities
Form semantic clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$
Compute semantic score u_m, semantic feature block $\mathbf{f}_{\text{sem}}^{m}$, and aggregated token features $\mathbf{f}_{\text{tok}}$
if $d =$ GATED **then**
  if $K \geq 2$ **then**
    $s \leftarrow u_m$
  else
    $s \leftarrow \mathcal{F}_G(\mathbf{f}_{\text{tok}})$
  end if
else if $d =$ STACKED **then**
  $\mathbf{f} \leftarrow [\mathbf{f}_{\text{sem}}^{m}, \mathbf{f}_{\text{tok}}]$
  $s \leftarrow \mathcal{F}_S(\mathbf{f})$
end if
return s

Higher values indicate greater uncertainty. When only one candidate is returned, $H_i^{(1)} = 0$, so the score retains only the across-response confidence-variation term.

A.3 CoCoA

CoCoA combines the confidence of a target response y^* with its semantic disagreement from N sampled alternatives [Vashurin et al., 2025]:

$$U_{\text{CoCoA}}(y^* \mid x) = u(y^* \mid x)\, \frac{1}{N} \sum_{i=1}^{N} \left[1 - s\Big(y^*, y^{(i)}\Big)\right],$$

where $s(\cdot, \cdot)$ is cosine similarity between response embeddings. We evaluate two variants: CoCoA SP uses sequence-level uncertainty, $u_{\text{SP}} = -\sum_{t=1}^{L} \log p(y_t^* \mid y_{<t}^*, x)$, whereas CoCoA PPL uses its length-normalised form, $u_{\text{PPL}} = -L^{-1} \sum_{t=1}^{L} \log p(y_t^* \mid y_{<t}^*, x)$. Neither uses supervised training labels, and both use the same sampled responses available to the other methods. Because SP sums log-probabilities over all L tokens while PPL divides by L, SP is sensitive to response length whereas PPL is not.

A.4 Gated and Stacked methods

Gated and Stacked combine semantic and token-level uncertainty in different ways. Gated uses a semantic score when the sampled responses form multiple clusters and a classifier over aggregated token features when they collapse into one cluster. Stacked removes this routing rule and learns jointly from the semantic representation, cluster count, and aggregated token features for every query. Algorithm 1 summarises both procedures.

B Additional experimental details

This section records the implementation choices shared across the evaluated methods, including the embedding model, clustering configuration, and supervised-classifier settings. Unless otherwise stated, these settings are fixed within each dataset and applied consistently across method variants.

Response embeddings. We embed every sampled response using `text-embedding-3-large` and compute cosine similarity between the resulting 3,072-dimensional vectors. No dimensionality reduction is applied before semantic clustering or construction of the response-similarity matrix.

Semantic clustering. The reported results use greedy representative-based clustering. Responses are processed in sampling order and assigned to the first existing cluster whose representative has cosine similarity at least τ; otherwise, a new cluster is created. The dataset-specific thresholds in Table 2 are selected on held-out validation data and then fixed before cross-validation. UEigV uses a Laplacian-eigenvalue threshold of $\epsilon = 0.1$, whereas the Von Neumann and spectral measures operate directly on the continuous similarity matrix.

Supervised classifiers. Both supervised methods use L_2-regularised logistic regression with `max_iter` $= 1000$. Gated standardises the aggregated token features and fits its classifier only on single-cluster training examples. Stacked standardises the combined semantic and token feature vector, reduces it to at most 15 principal components, and then fits the classifier. Scaling, principal-component analysis, and logistic regression are fitted within each training fold to prevent information leakage. The dataset-specific regularisation strengths C are reported in Table 2.

C Datasets, features, and labelling resources

This section provides the supporting details required to reproduce the evaluation data and labels. We first document the public and custom datasets, then list the token-level features used by the supervised methods, and finally provide the dataset-construction and hallucination-judging prompts.

C.1 Public datasets

Our public benchmarks cover domain knowledge, ambiguous questions, multi-hop reasoning, extractive question answering, and multimodal text recognition. They comprise AA Omni Finance, AmbigQA, HotpotQA, SQuAD, and Cheque Generation.

1. **AA Omni Finance** [Team, 2025]: the finance and law subset ($\sim$200 questions) of the AA-Omniscience cross-domain knowledge evaluation benchmark ($\sim$6,000 questions total, Apache-2.0), which measures factual accuracy of LLMs across diverse domains. Each question includes domain, topic, ground-truth answer, and metadata; answers are graded on a four-point scale (Correct / Incorrect / Partial / Not Attempted).
2. **AmbigQA** [Min et al., 2020]: derived from Google Natural Questions [Kwiatkowski et al., 2019], this dataset ($\sim$12,000 questions, CC BY-SA 3.0) targets inherently ambiguous open-domain questions that admit multiple valid answers depending on interpretation. Crowdworkers annotated each question with disambiguated question–answer pairs; we retain the original ambiguous question and evaluate against all plausible gold answers.
3. **HotpotQA** [Yang et al., 2018]: a multi-hop QA dataset ($\sim$7,400 dev questions, CC BY-SA 4.0) where answering requires reasoning over multiple Wikipedia paragraphs. No single paragraph contains the full answer. Sentence-level supporting facts are provided for explainability.
4. **Cheque Generation** [Verma, 2024]: a visual question answering dataset of handwritten cheque images (Apache-2.0) containing ground-truth Courtesy Amount (numeric) and Legal Amount (written-out) fields. Models receive cheque images as base64-encoded multimodal inputs and must perform precise numeric extraction, testing hallucination detection in vision-language settings.
5. **SQuAD** [Rajpurkar et al., 2016]: the Stanford Question Answering Dataset (100,000+ questions over 500+ Wikipedia articles, CC BY-SA 4.0), an extractive reading comprehension benchmark where the answer is a contiguous span within the source paragraph. It provides a context-grounded baseline where the answer is explicitly present in the input.

C.2 Custom datasets

We additionally evaluate Financial Summaries and Long-Text QA, two custom benchmarks designed to test hallucinations in finance-related summarisation and document-grounded question answering. Their construction procedures and prompts are reported in Appendix C.5.

1. **Financial Summaries** (finance domain-specific, 200 sampled, n=171 with a valid hallucination label): an LLM-generated summarisation dataset of financial texts covering AML

enforcement actions, cross-border M&A activity, and regulatory proceedings. Passages are crafted with deliberately confusable near-duplicate entity names and similar numeric figures to probe whether models introduce cross-contamination errors during summarisation.

2. **Long-Text QA** (finance domain-specific, n=150 sub-questions, n=30 question sets used for evaluation): 30 multi-part question sets (5 sub-questions each) over regulatory and financial filings, with one deliberately unanswerable sub-question per set to provide a controlled hallucination signal. It tests whether models fabricate answers when the source document does not contain the required information.

Table 4 summarises every dataset used in this work together with its license, source URL, sample size, and the modifications we applied.

Table 4: Datasets used in this paper: licenses, sources, and modifications. n here is the raw number of items sampled/generated before hallucination labelling; Table 2 reports the number with a valid label actually used in cross-validation (lower for AmbigQA, HotpotQA, Cheque Generation, and Financial Summaries, where judging failed for some queries).

Dataset	License	n	Source	Modifications
AmbigQA Min et al. [2020]	CC BY-SA 3.0	200	`https://huggingface.co/datasets/sewon/ambig_qa`	Randomly sampled 200 questions; reformatted into `Question/Answer` columns.
SQuAD Rajpurkar et al. [2016]	CC BY-SA 4.0	200	`https://huggingface.co/datasets/rajpurkar/squad`	Sampled 200 questions; reformatted with context passages.
HotpotQA Yang et al. [2018]	CC BY-SA 4.0	200	`https://huggingface.co/datasets/hotpotqa/hotpot_qa`	Sampled 200 questions from the dev split; reformatted.
AA Omni Finance Team [2025]	Apache 2.0	$\sim$200	`https://huggingface.co/datasets/ArtificialAnalysis/AA-Omniscience-Public`	Extracted finance subset; wrapped with domain-specific prompt.
Cheque Generation Verma [2024]	Apache 2.0	200	`https://huggingface.co/datasets/aniketVerma07/handwritten_cheque_vqa_dataset`	Sampled 200 cheque images; sent as base64-encoded inputs.
Financial Summaries	Internal	200	newly generated	LLM-synthesised financial passages with deliberate confusables.
Long-Text QA	Internal	150	newly generated	30 multi-part question sets over regulatory/financial documents.

C.3 Token-level feature catalogue

Table 5 lists the features extracted from each generated response. These measurements capture sequence confidence, token uncertainty, candidate margins, positional changes, and response length. Query-level token representations are formed by aggregating these features across the N sampled responses; the exact subsets used by Gated and Stacked are specified in Appendix C.4.

C.4 Exact classifier input features

Since our code is not released under an unrestricted public licence because of institutional data-governance constraints, we specify the exact feature subsets fed to each trained component here rather than leaving them implicit.

Gated cascade, $K = 1$ branch. Features are aggregated across the N sampled responses as described in §2.2, standardised, and passed to an L2-regularised logistic regression:

```
mean_max_topk_entropy, mean_max_abs_ent_delta, mean_p95_topk_entropy,
mean_min_margin, mean_length_norm_seqlogprob, confidence_spread,
mean_perplexity, std_mean_topk_entropy, mean_mid_mean_topk_entropy,
```

Table 5: Complete per-response token-level features. Features marked $\star$ are highlighted in §2.2.

Feature	Source	Description
Token log-probability features		
mean_logprob	$\ell_t^{(1)}$	Mean chosen-token log-prob
min_logprob	$\ell_t^{(1)}$	Min chosen-token log-prob
max_logprob	$\ell_t^{(1)}$	Max chosen-token log-prob
std_logprob	$\ell_t^{(1)}$	Std of chosen-token log-probs
p90_logprob	$\ell_t^{(1)}$	10th percentile (= 90th uncertainty pctl)
p95_logprob	$\ell_t^{(1)}$	5th percentile (= 95th uncertainty pctl)
logprob_variance	$\ell_t^{(1)}$	Variance of chosen-token log-probs
perplexity*	$\ell_t^{(1)}$	$\exp(-\frac{1}{T}\sum_t \ell_t^{(1)})$
length_norm_seqlogprob	$\ell_t^{(1)}$	Length-normalised sequence log-prob
seq_length	N/A	Number of tokens T_i
last1_logprob	$\ell_t^{(1)}$	Log-prob of last token
last3_mean_logprob	$\ell_t^{(1)}$	Mean log-prob of last 3 tokens
mean_lp_delta	$\Delta\ell_t$	Mean $\lvert\ell_{t+1}^{(1)} - \ell_t^{(1)}\rvert$
max_abs_lp_delta	$\Delta\ell_t$	Max $\lvert\ell_{t+1}^{(1)} - \ell_t^{(1)}\rvert$
std_lp_delta	$\Delta\ell_t$	Std of log-prob deltas
Top-k features (require top-k log-probs)		
mean_topk_entropy*	$H_{\text{topk}}(t)$	Mean top-k entropy (§2.1)
max_topk_entropy	$H_{\text{topk}}(t)$	Max top-k entropy
min_topk_entropy	$H_{\text{topk}}(t)$	Min top-k entropy
std_topk_entropy	$H_{\text{topk}}(t)$	Std of top-k entropy
p90_topk_entropy	$H_{\text{topk}}(t)$	90th percentile of top-k entropy
p95_topk_entropy	$H_{\text{topk}}(t)$	95th percentile of top-k entropy
last1_topk_entropy	$H_{\text{topk}}(t)$	Top-k entropy of last token
last3_mean_topk_entropy	$H_{\text{topk}}(t)$	Mean top-k entropy of last 3 tokens
mean_ent_delta	ΔH_t	Mean $\lvert H_{\text{topk}}(t+1) - H_{\text{topk}}(t)\rvert$
max_abs_ent_delta	ΔH_t	Max entropy delta (spike signal)
std_ent_delta	ΔH_t	Std of entropy deltas
mean_entropy_alts*	ranks 2–k	Mean entropy over non-greedy tokens
max_entropy_alts	ranks 2–k	Max entropy over non-greedy tokens
mean_margin*	$\ell_t^{(1-2)}$	Mean $(\ell_t^{(1)} - \ell_t^{(2)})$
min_margin	$\ell_t^{(1-2)}$	Min margin (least decisive position)
std_margin	$\ell_t^{(1-2)}$	Std of margins
p10_margin	$\ell_t^{(1-2)}$	10th percentile margin
frac_non_greedy	top-k	Fraction of positions where chosen $\neq$ argmax

```
mean_end_mean_topk_entropy, mean_mid_max_ent_spike,
entropy_drift_start_to_end, mid_vs_start_spike_ratio,
margin_decay_start_to_end
```

When top-k log-probabilities are unavailable (`has_topk=False`, e.g. some model and API configurations; see Appendix E.4), the following reduced, chosen-token-only feature set is used instead:

```
mean_max_logprob, mean_max_abs_lp_delta, mean_p95_logprob,
mean_min_logprob, mean_length_norm_seqlogprob, confidence_spread,
mean_perplexity, std_mean_logprob, mean_mid_mean_logprob,
mean_end_mean_logprob, entropy_drift_start_to_end,
margin_decay_start_to_end
```

Stacked classifier. This model pools one SE-variant feature block with the full aggregated token feature vector from Table 5. It then applies StandardScaler, PCA with up to 15 components, and L2-regularised logistic regression. The SE-variant block is one of:

```
Hybrid:  se_hybrid, s_hat_hybrid, n_clusters
Spectral:  spectral_total, spectral_erank, spectral_epistemic,
```

```
n_clusters
Von Neumann:  se_von_neumann, n_clusters
```

C.5 Dataset construction

The following materials document the construction of the two custom benchmarks. We retain the complete instructions and question sets to make the evaluation design reproducible.

C.5.1 Long-Text QA

The Long-Text QA dataset comprises 30 multi-part question sets (5 sub-questions each, 150 sub-questions total) over three regulatory and financial source documents: *SR 11-7: Guidance on Model Risk Management* (Federal Reserve / OCC), an anonymised 10-K annual filing, and the *EU AI Act*. Each question set is presented to the model together with the full source document as context, using the shared prompt template below. Exactly one sub-question per set (always Q5) is *deliberately unanswerable* from the source document; the gold answer for that question is "Not provided in context." This design provides a controlled hallucination signal: a model that fabricates an answer to Q5 produces a detectable hallucination.

Prompt template. All 30 question sets share the following instruction prefix (question content varies per set):

```
Answer the questions based on the document attached.  Use only the
context provided in the document and do not use your training data
for this.  There are five questions.  Return the output in the
provided output format.  Strictly follow the output format.  Do
not add anything else.  If you do not know the answer or the answer
is not present in the attached document, output ‘‘Not provided in
context.’’

INPUT Format:  ['question 1', 'question 2', ..., 'question 5']

OUTPUT format:
{'question 1':  'answer', 'question 2':  'answer', ..., 'question 5':
'answer'}
```

Complete question listing. Table 6 lists all 30 question sets. The unanswerable sub-question (Q5) in each set is marked with †.

Table 6: **Long-Text QA: all 30 question sets.** Each set contains five sub-questions posed over the indicated source document. Q5 (marked †) is deliberately unanswerable from the source; the expected answer is "Not provided in context."

Set	Source	Sub-questions
1	SR 11-7	Q1: Explain the core elements of an effective validation framework. Q2: Explain the processes and activities associated with model validation. Q3: What is the definition of a Model according to SR 11-7? Q4: What is the exact definition of model validation according to SR 11-7? Q5†: What are the five key elements of comprehensive validation?
2	SR 11-7	Q1: What should a bank's board of directors and senior management demand from model validation? Q2: What are the three core elements of effective model risk management? Q3: What are the key components of a robust model validation framework? Q4: What types of outcomes analysis should be performed during model validation? Q5†: What are the regulatory penalties for non-compliance with SR 11-7?

Continued on next page

Set	Source	Sub-questions
3	SR 11-7	Q1: What is the role of conceptual soundness evaluation in model validation? Q2: What does SR 11-7 say about the use of vendor models? Q3: How should model limitations be documented according to SR 11-7? Q4: What are the key elements of an effective model inventory? Q5†: What specific software tools does SR 11-7 recommend for model validation?
4	SR 11-7	Q1: What is the definition of model risk according to SR 11-7? Q2: What are the two main sources of model risk identified in SR 11-7? Q3: Why might models with good performance become inaccurate over time? Q4: How does SR 11-7 define an effective challenge? Q5†: What penalties can be imposed for failing to comply with SR 11-7?
5	SR 11-7	Q1: What is the suggested frequency of model validation reviews? Q2: What role does internal audit play in model risk management? Q3: How should banks manage model risk for models developed by vendors or third parties? Q4: What are the key considerations for using sensitivity analysis in model validation? Q5†: What machine learning methods does SR 11-7 recommend for model validation?
6	SR 11-7	Q1: What is the role of the board of directors in model risk management according to SR 11-7? Q2: What does SR 11-7 say about the documentation requirements for models? Q3: How should banks handle models that show declining performance? Q4: What are the key elements of an effective model development process? Q5†: What cloud computing requirements does SR 11-7 specify?
7	SR 11-7	Q1: What does SR 11-7 say about model governance policies? Q2: How should banks assess model limitations? Q3: What is the role of outcomes analysis in model validation? Q4: What are the key considerations for back-testing models? Q5†: What specific stress testing scenarios does SR 11-7 require banks to run?
8	10-K	Q1: How many people does [ANON] employ globally? Give time as well. Q2: What are the MD&A-risk factors? What happens if they are not adhered to? Q3: Summarise the Description of Securities Registered Pursuant to Section 12 of the Securities Exchange Act of 1934 as provided in Exhibit 4.2. Q4: Describe the [X.XXX]% Fixed-to-Floating Rate Preferred Capital Securities. Q5†: Describe the deliquidification process?
9	10-K	Q1: What is the position of [EXEC_A]? Q2: What is the position of [EXEC_B]? Q3: What does the Vesting Schedule indicate? Q4: What does Risk Adjustment Process indicate? Q5†: What is the position of [EXEC_C]?
10	10-K	Q1: Is the registration statement [REG_NO_1] included? Q2: Is the registration statement [REG_NO_2] included? Q3: What does TSR mean? Q4: What does Transition Period mean? Q5†: What does TSP mean?

Continued on next page

Set	Source	Sub-questions
11	10-K	Q1: What is the definition of Firm in this document? Q2: What Financial Reporting Measures are used? Q3: What is the par value of the common stock listed under securities registered? Q4: How much do the preferred securities initially pay? Q5†: How does the vesting schedule differ from the non-vesting schedule?
12	10-K	Q1: Who is the issuing entity? Q2: Are there any Securities registered pursuant to Section 12(g) of the Act? Q3: Who are the current Executive Officers? Q4: How many Executive Officers are there at [ANON] currently? Q5†: How many shares does the company need to issue as per the reinstated certificate of incorporation?
13	10-K	Q1: What is the main banking subsidiary of [ANON] in continental Europe? Q2: At the end of 2023, what % of [ANON]'s global workforce were women? Q3: At the end of 2023, what % of [ANON]'s executive committee were women? Q4: At Dec. 31, 2023, approximately what % of [ANON]'s total employees were based outside the U.S.? Q5†: At Dec. 31, 2023, approximately what % of [ANON]'s total employees were based in India?
14	10-K	Q1: Where are the corporate headquarters of [ANON] located? Q2: Who supervised the evaluation of the effectiveness of [ANON]'s disclosure controls and procedures? Q3: Who is the oldest Executive Officer? Q4: Do the holders of Common Stock have cumulative voting rights? Q5†: Do the holders of Dispersed Stock have cumulative voting rights?
15	10-K	Q1: Who has right to request that the Secretary call a special meeting of stockholders? Q2: Approximately how many participants were covered by the frozen U.S. defined benefit pension plan at December 31, 2023? Q3: Approximately how many participants were covered by non-U.S. defined benefit plans at December 31, 2023? Q4: What is the approximate size of the company's headquarters building? Q5†: What is the approximate size of the company's office building in [CITY_A]?
16	10-K	Q1: What is included in the Other segment? Q2: What were assets under custody and/or administration at December 31, 2023? Q3: What were assets under management at December 31, 2023? Q4: Which principal U.S. banking subsidiary houses [SEGMENT_A] businesses? Q5†: Who is the Chief AI Officer of [ANON]?
17	10-K	Q1: How many employees were in EMEA at the end of 2023? Q2: How many employees were in APAC at the end of 2023? Q3: What percentage of the global workforce were women at the end of 2023? Q4: What percentage of the U.S. workforce were from underrepresented ethnic and/or racial backgrounds at the end of 2023? Q5†: How many employees were in APAC at the end of 2025?

Continued on next page

Set	Source	Sub-questions
18	10-K	Q1: What equity award was given to eligible employees as described in Human Capital Management? Q2: How many participants were in the 401(k) plan at December 31, 2023? Q3: How many U.S. participants were covered by the frozen defined benefit pension plan at December 31, 2023? Q4: How many non-U.S. participants were covered by non-U.S. defined benefit plans at December 31, 2023? Q5†: How many participants were in the dental insurance plan at December 31, 2023?
19	10-K	Q1: How much space does the company have in EMEA? Q2: How much leased space does the company have in APAC? Q3: How many holders of record of common stock were there as of January 31, 2024? Q4: Were the company's disclosure controls and procedures effective as of December 31, 2023? Q5†: How much leased space does the company have in Germany?
20	10-K	Q1: What PCAOB firm identification number is listed for [AUDIT_FIRM]? Q2: What was fee and other revenue in 2023? Q3: What was net interest revenue in 2023? Q4: What was total revenue in 2023? Q5†: What was total revenue in 2003?
21	10-K	Q1: What is the registrant's legal name in the filing? Q2: What is the company's stock ticker symbol? Q3: On which exchange is the common stock listed? Q4: What was the aggregate market value of common stock held by non-affiliates as of June 30, 2023? Q5†: What is the company's stock ticker symbol for NSE?
22	10-K	Q1: What was net income applicable to common shareholders in 2023? Q2: What were diluted earnings per common share in 2023? Q3: What was return on common equity in 2023? Q4: What was return on tangible common equity in 2023? Q5†: What was return on tangible common equity in 2003?
23	10-K	Q1: What was the common cash dividend per share in 2023? Q2: What was the common dividend payout ratio in 2023? Q3: What was the closing stock price at December 31, 2023? Q4: What was market capitalisation at December 31, 2023? Q5†: What was market capitalisation at December 31, 2003?
24	10-K	Q1: What was book value per common share at December 31, 2023? Q2: What was tangible book value per common share at December 31, 2023? Q3: What was the CET1 ratio at December 31, 2023? Q4: What was the Tier 1 capital ratio at December 31, 2023? Q5†: What was the Tier 1 capital ratio at December 31, 2003?
25	10-K	Q1: What was the total capital ratio at December 31, 2023? Q2: What was the Tier 1 leverage ratio at December 31, 2023? Q3: What was the supplementary leverage ratio at December 31, 2023? Q4: What was the effective tax rate in 2023? Q5†: What was the effective tax rate in 2003?

Continued on next page

Set	Source	Sub-questions
26	10-K	Q1: What was the FDIC special assessment accrual recorded in 2023 noninterest expense? Q2: What adjustment was made in February 2024 related to the FDIC special assessment? Q3: What was the average common shares outstanding on a diluted basis in 2023? Q4: What Financial Reporting Measures are used under the Recovery of Erroneously Awarded Incentive-Based Compensation Policy? Q5†: What was the average common shares outstanding on a diluted basis in 2003?
27	EU AI Act	Q1: What is the main purpose of the EU AI Act? Q2: What kinds of rules does the EU AI Act lay down? Q3: To whom does the EU AI Act apply? Q4: Does the EU AI Act apply to AI used exclusively for military, defence or national security purposes? Q5†: What special provisions are made for AI in financial services?
28	EU AI Act	Q1: Does the EU AI Act apply to AI systems developed solely for scientific research and development? Q2: What is an 'AI system' under the EU AI Act? Q3: What does the EU AI Act mean by 'provider'? Q4: What is a 'deployer' under the Act? Q5†: What is meant by 'subscriber'?
29	EU AI Act	Q1: What does the EU AI Act mean by 'provider'? Q2: What is a 'deployer' under the Act? Q3: What is meant by 'intended purpose' in the EU AI Act? Q4: What is 'reasonably foreseeable misuse'? Q5†: Is the Act applicable to the United Kingdom of Great Britain and Ireland?
30	EU AI Act	Q1: What is 'AI literacy' under the Act? Q2: What obligation does the EU AI Act impose regarding AI literacy? Q3: What overall regulatory approach does the EU AI Act use? Q4: What AI practice involving subliminal or manipulative techniques is prohibited? Q5†: What exemptions are provided for BigTech companies under the Act?

C.5.2 Financial Summaries

The following prompt template is used to generate the synthetic finance texts in our summarisation hallucination dataset (Section 3.2). Placeholders in braces are filled programmatically for each sample.

```
You are constructing a research dataset for studying LLM
hallucination in financial text summarisation. Your task is to
generate a synthetic finance text that is designed to trigger a
specific type of hallucination when an LLM attempts to summarise it.

HALLUCINATION TYPE: {type_code} - {type_name}
TYPE DESCRIPTION: {type_description}

{trigger_instructions}

FINANCE SUB-DOMAIN: {subdomain}

CONSTRAINTS:
 1. The text must be approximately 500 words (450-550 acceptable
    range).
 2. The text must read like a realistic finance document (news
    article, analyst note, regulatory filing excerpt, earnings
    summary, etc.) in the specified sub-domain.
```

```
  3. All entities, figures, and events in the text are FICTIONAL but
     must be plausible. Do not use real company names, real people,
     or real events.
  4. The text must be self-contained (no references to external
     documents).
  5. Do NOT include any meta-commentary about the hallucination
     design.  The text should look completely natural.

VARIATION SEED: {variation_seed}
Use this seed to ensure this text is distinct from others in the
same category. Vary the entities, scenario, writing style, and
structure.

Return your response as a JSON object with these exact fields:
{
    "text":  "<the ~500-word finance text>",
    "ground_truth_summary":  "<a correct 3-4 sentence summary that
faithfully captures the key facts without any distortion>",
    "key_facts":  ["<fact 1>", "<fact 2>", ...],
    "hallucination_traps":  ["<description of trap 1>",
"<description of trap 2>", ...],
    "design_notes":  "<brief explanation of what structural features
were embedded>"
}

key_facts:  An exhaustive list of every verifiable claim in the text
(aim for 8-15 facts).

hallucination_traps:  Specific descriptions of what a summariser
is likely to get wrong (e.g. "May attribute Meridian's $2.1B
revenue to Apex Corp").  Aim for 3-5 traps. For CTRL texts, set
hallucination_traps to an empty list.

Return ONLY the JSON object, no other text.
```

Per-type trigger instructions The `{trigger_instructions}` placeholder in the prompt above is filled with one of the following type-specific structural requirement blocks. Each block steers the generator toward text structures that are known to elicit the corresponding hallucination type during summarisation.

ES – Entity Swap.

```
STRUCTURAL REQUIREMENTS for Entity Swap triggers:
  1. Include at least 3 named entities of the same type (e.g. 3 banks, 3
     CEOs, 3 tickers) in close proximity within the text.
  2. Interleave metrics and attributes across entities so that a careless
     reader (or LLM) could easily attribute Entity A's figures to
     Entity B.
  3. Use similar-sounding or related entity names (e.g. Apex Securities
     and Meridian Capital, or Northern Bank and Pacific Trust) to increase
     confusion potential.
  4. Ensure each entity has distinct, verifiable facts that must not be
     swapped.
```

NE – Numerical Error.

```
STRUCTURAL REQUIREMENTS for Numerical Error triggers:
  1. Pack the text with at least 8-10 specific numbers:  revenue figures,
     percentages, basis points, dates, headcounts, or dollar amounts.
  2. Include numbers that differ by small amounts (e.g. 3.2% vs 3.7%,
     $14.2B vs $14.8B).
  3. Mix absolute values with percentages and year-over-year changes in
     the same paragraph.
  4. Include at least one number that contradicts an intuitive
     expectation.
```

NF – Negation Flip.

```
STRUCTURAL REQUIREMENTS for Negation Flip triggers:
  1. Include at least 2-3 negated statements that carry critical meaning
     (e.g. ''the board did NOT approve'', ''excluding derivatives exposure'').
  2. Bury negations in subordinate clauses or complex sentence structures.
```

3. Use double negatives, ‘‘except’’/‘‘excluding’’/‘‘unless’’ constructions.
4. Place an affirmative statement near a negated one about a related topic so a summariser might conflate them.

TC – Temporal Confusion. STRUCTURAL REQUIREMENTS for Temporal Confusion triggers:

1. Present events in non-chronological order (e.g. mention Q4 results before Q2 events).
2. Include at least 4-5 distinct date references (specific months, quarters, years).
3. Describe actions that were planned, deferred, completed, or reversed across different time periods.
4. Use ambiguous temporal language (‘‘previously’’, ‘‘earlier this year’’, ‘‘following the announcement’’) alongside specific dates.

CF – Causal Fabrication. STRUCTURAL REQUIREMENTS for Causal Fabrication triggers:

1. Juxtapose two events or facts that are temporally or thematically related but have NO stated causal connection.
2. A summariser is likely to infer ‘‘A caused B’’ or ‘‘A led to B’’ when the text only says ‘‘A happened. Separately, B happened.’’
3. Include at least 2 such juxtaposition pairs.
4. Explicitly avoid causal language (‘‘because’’, ‘‘therefore’’, ‘‘as a result’’) between the juxtaposed events.

UI – Unsupported Inference. STRUCTURAL REQUIREMENTS for Unsupported Inference triggers:

1. Present raw data, metrics, or observations WITHOUT drawing conclusions.
2. A summariser is likely to add interpretive statements like ‘‘this suggests’’, ‘‘indicating strong performance’’, or ‘‘raising concerns about’’.
3. Include data that could support multiple contradictory interpretations.
4. Avoid any evaluative or interpretive language in the source text.

DF – Detail Fabrication. STRUCTURAL REQUIREMENTS for Detail Fabrication triggers:

1. Leave deliberate gaps: mention ‘‘a major US bank’’ without naming it, refer to ‘‘the acquiring firm’’ without specifics, cite ‘‘an undisclosed sum’’.
2. Include partial information that a summariser might ‘‘complete’’ by fabricating plausible details (names, amounts, locations).
3. Reference unnamed parties, unspecified contract terms, or redacted figures.
4. Include at least 3 such deliberate vagueness points.

MC – Merging Claims. STRUCTURAL REQUIREMENTS for Merging Claims triggers:

1. Include at least 2 pairs of related but distinct claims in adjacent sentences.
2. For example: ‘‘Bank A expanded its equities desk in London’’ followed by ‘‘Bank A also opened a new fixed-income operation in Frankfurt’’ - a summariser might merge these into ‘‘Bank A expanded its equities and fixed-income operations in London and Frankfurt’’ (incorrectly linking equities to Frankfurt).
3. Make the claims similar enough to invite merging but with distinct specifics.

OQ – Omission of Qualifiers. STRUCTURAL REQUIREMENTS for Omission of Qualifiers triggers:

1. Saturate the text with hedging language: ‘‘preliminary estimates suggest’’, ‘‘subject to regulatory approval’’, ‘‘under certain market conditions’’, ‘‘the committee may consider’’, ‘‘pending final review’’.
2. At least 4-5 key claims should be heavily qualified.
3. Make the underlying claims significant enough that dropping the qualifier materially changes the meaning (e.g. ‘‘may acquire’’ vs ‘‘will acquire’’).

OG – Over-generalisation.

```
STRUCTURAL REQUIREMENTS for Over-generalisation triggers:
1. Make every claim narrowly scoped: specify region, time period,
   client segment, product line, or division.
2. For example: ''Among retail clients in the APAC region during Q3
   2025, adoption of digital advisory tools rose 12%.'' A summariser
   might drop the qualifiers and report ''digital advisory adoption rose
   12%'' as if it were global/universal.
3. Include at least 3-4 such narrowly scoped claims.
```

CTRL – Control (easy).

```
STRUCTURAL REQUIREMENTS for Control (easy) texts:
1. Write a clear, well-structured finance text with unambiguous facts.
2. Use straightforward sentence structure with no buried negations or
   tricky phrasing.
3. Include a reasonable number of facts (4-6) that are easy to verify in
   a summary.
4. Avoid any adversarial structural patterns. This text should be easy
   to summarise correctly.
```

C.6 LLM judge prompts for hallucination labels

We use GPT-4.1 for four judging protocols according to the reference information available for each task. The blind judge evaluates an answer without a gold reference, the binary and scalar judges compare it with an expected answer, and the summarisation judge checks generated claims against the source text. The complete prompts are reproduced below.

C.6.1 Blind judge

The blind judge receives only the question and the model's answer; no reference material is provided. It returns a continuous hallucination probability, a confidence score, and free-text reasoning. The prompt instructs the judge to look for suspiciously specific details, contradiction of widely known facts, false confidence, internal inconsistency, and confabulation.

```
You are an expert evaluator. Given a question and the model's
answer, assess whether the answer is likely to contain hallucinated
or fabricated information.

You do NOT have access to the correct answer or any reference
material. Use only your knowledge and reasoning to judge the
answer.

Look for the following signals of hallucination:
 1. Suspiciously specific details (e.g. invented names, dates,
    statistics)
 2. Contradiction of widely known facts
 3. False confidence about uncertain or obscure topics
 4. Internal inconsistency within the answer
 5. Confabulation: plausible-sounding but fabricated information

Return: {"hallucination_probability": <0.0-1.0>, "confidence":
<0.0-1.0>, "reasoning": "..."}
```

C.6.2 Binary judge

The binary judge compares the proposed answer against an expected answer and returns a binary verdict (1 = correct, 0 = incorrect). Inspired by Kuhn et al. [2023] and Farquhar et al. [2024], the prompt explicitly permits additional detail beyond the expected answer while requiring that the core claim not be wrong.

```
We are assessing the quality of answers to the following question.

QUESTION: {question}
EXPECTED ANSWER: {expected_answer}
PROPOSED ANSWER: {proposed_answer}
```

```
Important: It is acceptable for the proposed answer to include
additional details beyond the expected answer. The core of the
proposed answer must not be incorrect, it may be incomplete, but it
must not be wrong.

Within the context of the question, does the proposed answer mean
the same as the expected answer?

Respond ONLY with "yes" or "no".
```

C.6.3 Scalar judge

The scalar judge uses the same inputs as the binary judge but returns a continuous consistency score on a 0–100 scale together with free-text reasoning.

```
Rate the level of consistency between the proposed answer and the
expected answer to the question, from 0 to 100.

Important: It is acceptable for the proposed answer to include
additional details beyond the expected answer. The core of the
proposed answer must not be incorrect-it may be incomplete, but it
must not be wrong. A correct but incomplete answer should still
receive a high score.

Question: {question}
Expected answer: {expected_answer}
Proposed answer: {proposed_answer}

Return: {"rating": <0-100>, "reasoning": "..."}
```

C.6.4 Summarisation judge

The summarisation judge evaluates whether an LLM-generated summary of a financial text contains hallucinations. It receives the source text, a list of key facts extracted from the source, and the summary to evaluate. The prompt defines a ten-category hallucination taxonomy—Entity Swap (ES), Numerical Error (NE), Negation Flip (NF), Temporal Confusion (TC), Causal Fabrication (CF), Unsupported Inference (UI), Detail Fabrication (DF), Merging Claims (MC), Omission of Qualifiers (OQ), and Over-generalisation (OG)—and asks the judge to return structured JSON with a binary flag, severity rating, and per-claim explanations. In practice, two independent GPT-4.1 judgements are combined using the union verdict.

```
You are an expert evaluator assessing whether an LLM-generated
summary of a financial text contains hallucinations. A
hallucination is any claim in the summary that is not supported
by, or contradicts, the source text.

HALLUCINATION TAXONOMY:
- ES (Entity Swap): Attributes facts to the wrong entity
- NE (Numerical Error): Incorrect numbers, percentages, or dates
- NF (Negation Flip): Inverts or drops a negation
- TC (Temporal Confusion): Wrong ordering or dating of events
- CF (Causal Fabrication): Invents cause-effect links not in the
source
- UI (Unsupported Inference): Draws conclusions beyond the text
- DF (Detail Fabrication): Invents specific details absent from the
source
- MC (Merging Claims): Fuses distinct statements into a false
composite
- OQ (Omission of Qualifiers): Drops hedging, making tentative
claims definitive
- OG (Over-generalisation): Broadens a narrowly scoped claim

SOURCE TEXT:
{source_text}
```

```
KEY FACTS (exhaustive list of verifiable claims in the source):
{key_facts}

SUMMARY TO EVALUATE:
{summary}

Evaluate the summary carefully.  Return a JSON object:
{{
  "has_hallucination":  <true or false>,
  "hallucination_types_found":  ["<code>", ...],
  "hallucinated_claims":  [{{
    "claim":  "<the hallucinated statement from the summary>",
    "type":  "<hallucination type code>",
    "explanation":  "<why this is a hallucination, referencing the
source text>"
  }}],
  "severity":  "<none|low|medium|high>",
  "overall_reasoning":  "<2-3 sentence assessment>"
}}

If the summary is faithful, set has_hallucination to false,
hallucination_types_found and hallucinated_claims to empty lists,
severity to "none".

Return ONLY the JSON object.
```

D Extended diagnostic results

This section provides additional evidence supporting the main results and discussion. We first examine the complementary failure modes of semantic and token uncertainty, then report ROC curves and cross-model comparisons, including CoCoA.

D.1 Complementary failure modes

Figures 6 and 7 examine the principal blind spot of semantic entropy. Among hallucinated queries, the proportion whose N responses collapse into one semantic cluster ranges from 39% on AmbigQA to 99% on Financial Summaries, forcing Standard SE to zero. Within these single-cluster cases, median TopK uncertainty remains higher for hallucinated queries on six of seven datasets, showing that token probabilities often retain information after semantic disagreement disappears. The exception is SQuAD, and substantial distributional overlap remains elsewhere: between 21% and 56% of hallucinated single-cluster cases also fall below the corresponding non-hallucinated TopK median. Thus, the two signals are complementary, but neither guarantees detection when the model is both semantically consistent and token-confident.

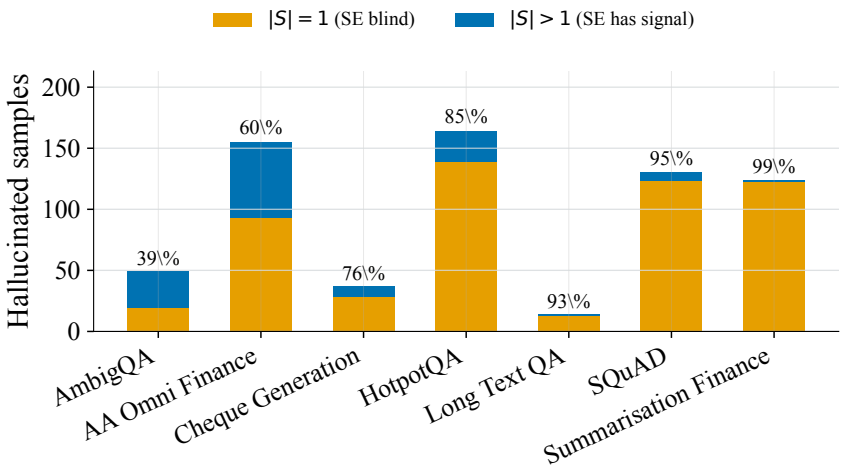

Figure 6: **Single-cluster collapse among hallucinated queries.** Hallucinated queries are divided according to whether their N sampled responses form one semantic cluster ($K = 1$) or multiple clusters ($K \geq 2$). Standard SE is necessarily zero in the single-cluster group. Percentages above the bars report the $K = 1$ share for each dataset.

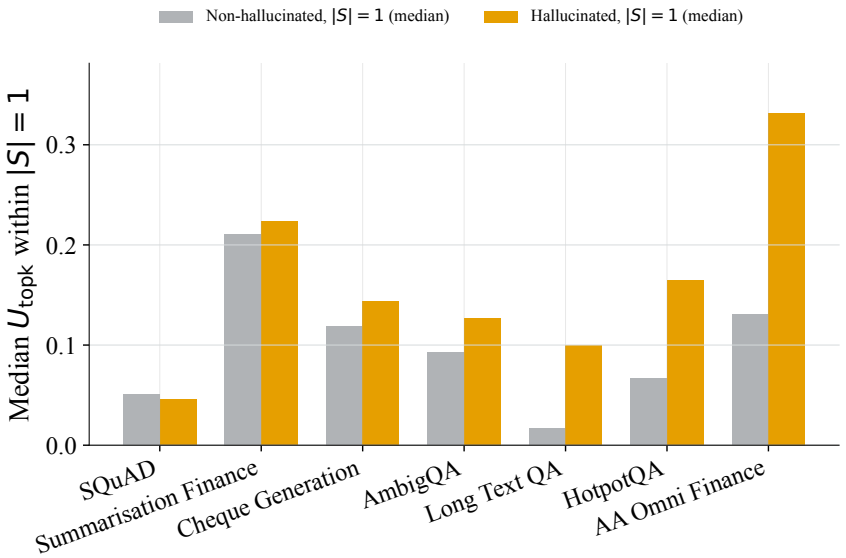

Figure 7: **TopK uncertainty within the semantic-entropy blind spot.** The analysis is restricted to $K = 1$ queries, for which Standard SE is zero. Bars compare median TopK uncertainty for hallucinated and non-hallucinated queries. The annotation for each dataset reports the proportion of hallucinated queries whose TopK score also falls below the non-hallucinated median.

D.2 Token-feature discriminability

Figure 8 relates the strongest token-level effect size on each dataset to the performance of the Stacked classifier. Datasets with clearer token-feature separation generally yield stronger Stacked performance, whereas Financial Summaries remains weak on both measures. This relationship is descriptive rather than causal: it is based on seven datasets, and Stacked also uses semantic features rather than the displayed token feature alone.

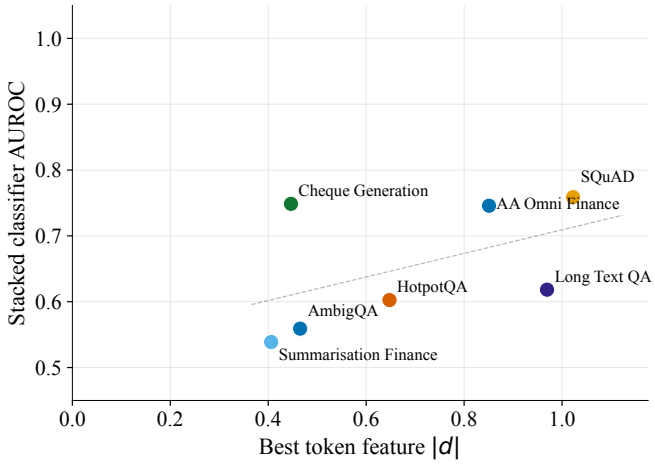

Figure 8: **Token-feature discriminability and Stacked performance.** Stacked AUROC plotted against the largest absolute Cohen's d among token-level features for each dataset. Stronger token-feature separation is generally associated with higher Stacked AUROC, while Financial Summaries is the clearest low-separation outlier.

D.3 ROC curves across datasets

Figure 9 shows how method sensitivity changes across false-positive-rate thresholds, including CoCoA SP and CoCoA PPL. The curves reinforce the dataset-dependent rankings reported in the main text: relative performance varies across both datasets and operating points.

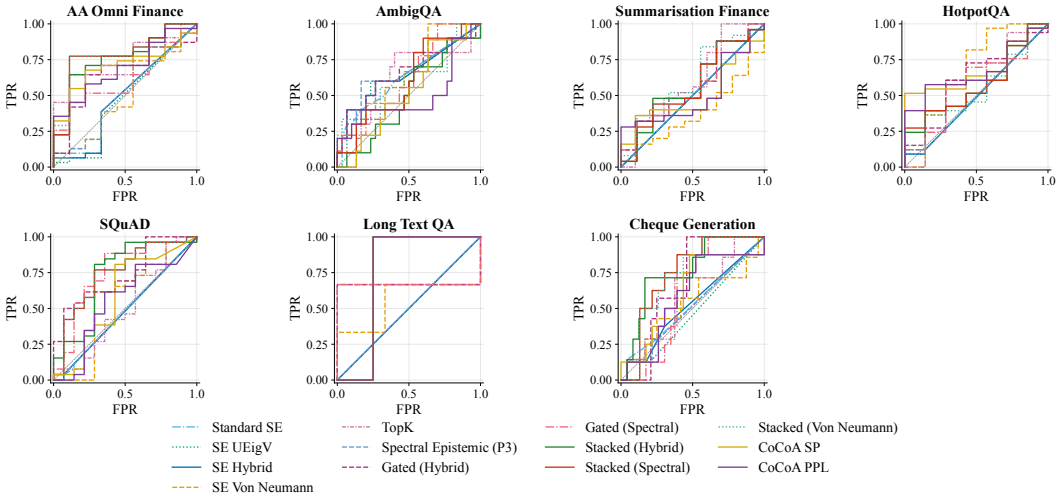

Figure 9: **ROC curves across datasets, with CoCoA.** Each panel shows the representative held-out-fold ROC curves from the five-fold evaluation of the canonical model run, including CoCoA SP and CoCoA PPL computed from the same generations; the diagonal denotes chance performance.

D.4 Cheque Generation robustness

Figure 10 compares method categories across response counts for the three vision-capable models. Increasing N generally improves the strongest semantic method for GPT-5.1 and GPT-5.4, but TopK, Gated, and Stacked vary comparatively little. Thus, additional responses do not produce a consistent gain for the combined methods. These matched sweeps were scored externally and are reported separately from the canonical pooled out-of-fold evaluation because their scoring pipeline could not be fully verified against our evaluator.

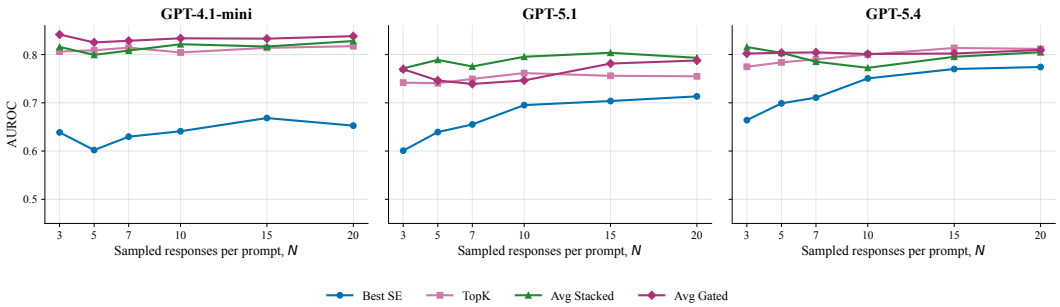

Figure 10: **Cheque Generation robustness across models and response counts.** AUROC is shown for the strongest semantic method, TopK, the mean of three Stacked variants, and the mean of two Gated variants over $N \in \{3, 5, 7, 10, 15, 20\}$. The externally scored sweeps are kept separate from the primary pooled out-of-fold results. CoCoA is omitted because this sweep contains no CoCoA scores.

E Ablation studies

This section examines sensitivity to model family, response count N, sampling temperature, the number of returned token candidates k, and classifier calibration. We report each factor separately to distinguish changes in the available uncertainty evidence from changes in the fitted method.

E.1 Model-family ablation

Table 7 reports the primary GPT-4.1-mini fold-mean results discussed in Section 4.1. Tables 8–11 report pooled out-of-fold results for the 23 available text model–dataset settings using GPT-5.1, GPT-5.4, Llama 3.3 70B, and a regenerated GPT-4.1-mini run. These tables use text-only agent configurations; Appendix D.4 separately reports vision-capable Cheque Generation settings for GPT-4.1-mini, GPT-5.1, and GPT-5.4, giving 26 available settings overall. CoCoA SP and CoCoA PPL are computed from the same generated responses as the other methods that do not use supervised training labels. The dataset-specific τ and C values are held fixed across models rather than retuned, and the single-class Llama Long-Text QA setting is marked as N/A.

Table 7: GPT-4.1-mini AUROC by method and dataset on the held-out folds of 5-fold cross-validation. Generation uses N=10, temperature 1.0, and five returned log-probability candidates per token (k=5). CoCoA SP and PPL use their respective confidence variants with semantic consistency computed from the same generations. Each value is the mean of the five fold-level test AUROCs. Bold marks the highest AUROC in each dataset.

Method	AA Fin.	AmbigQA	Summ. Fin.	HotpotQA	SQuAD	Long QA	Cheque
Standard SE	0.508	0.631	0.504	0.504	0.454	0.475	0.481
SE UEigV	0.469	0.618	0.504	0.500	0.454	0.475	0.498
SE Hybrid	0.473	0.618	0.504	0.500	0.455	0.475	0.501
SE Von Neumann	0.518	0.652	0.468	0.638	0.558	0.589	0.581
TopK	0.747	0.637	**0.611**	0.664	0.451	0.681	0.600
Spectral Epistemic	0.517	0.649	0.504	0.503	0.454	0.475	0.482
CoCoA SP	0.687	**0.667**	0.513	0.681	0.575	0.683	0.618
CoCoA PPL	0.704	0.654	0.516	**0.696**	0.546	0.661	0.609
Gated Hybrid	0.619	0.625	0.575	0.677	0.729	**0.731**	0.637
Gated Spectral	0.631	0.617	0.573	0.651	0.753	0.711	0.607
Stacked Hybrid	**0.763**	0.532	0.571	0.621	0.758	0.683	**0.760**
Stacked Spectral	0.756	0.560	0.569	0.628	**0.762**	0.683	0.740
Stacked Von Neumann	0.751	0.552	0.586	0.621	0.760	0.683	0.738

Pooled out-of-fold AUROC differs by 0.01 to 0.02 on most cells. It ranks TopK marginally above Gated Hybrid on HotpotQA and Long Text QA. Appendix F.2 reports pooled estimates and 95% bootstrap intervals for every cell.

Table 8: Complete pooled out-of-fold AUROC for GPT-4.1-mini-new. Missing generation, judging, or modality combinations are shown as N/A; Cheque Generation is reported separately because these agents are text-only.

Method	AA Finance	AmbigQA	HotpotQA	SQuAD	Long Text	Summ. Finance
Standard SE	0.522	0.634	0.540	0.442	0.562	0.500
SE UEigV	0.509	0.617	0.535	0.441	0.562	0.500
SE Hybrid	0.515	0.607	0.533	0.442	0.562	0.500
SE Von Neumann	0.561	0.623	0.676	0.524	0.771	0.442
TopK	0.621	0.589	0.687	0.457	0.729	0.309
Spectral Epistemic (P3)	0.534	0.655	0.537	0.444	0.562	0.500
CoCoA SP	0.621	0.661	0.671	0.562	0.771	0.327
CoCoA PPL	0.633	0.683	0.684	0.538	0.722	0.329
Gated (Hybrid)	0.507	0.581	0.636	0.735	0.521	0.240
Gated (Spectral)	0.494	0.575	0.593	0.753	0.493	0.240
Stacked (Hybrid)	0.634	0.448	0.653	0.749	0.604	0.524
Stacked (Spectral)	0.640	0.474	0.665	0.763	0.611	0.506
Stacked (Von Neumann)	0.636	0.451	0.674	0.767	0.597	0.520

Table 9: Complete pooled out-of-fold AUROC for Llama-70B. Missing generation, judging, or modality combinations are shown as N/A; Cheque Generation is reported separately because these agents are text-only.

Method	AA Finance	AmbigQA	HotpotQA	SQuAD	Long Text	Summ. Finance
Standard SE	0.515	0.503	0.555	0.523	N/A	0.487
SE UEigV	0.514	0.502	0.555	0.523	N/A	0.487
SE Hybrid	0.516	0.506	0.557	0.523	N/A	0.487
SE Von Neumann	0.501	0.538	0.569	0.479	N/A	0.306
TopK	0.584	0.593	0.650	0.541	N/A	0.416
Spectral Epistemic (P3)	0.513	0.507	0.556	0.523	N/A	0.487
CoCoA SP	0.533	0.654	0.642	0.598	N/A	0.325
CoCoA PPL	0.542	0.632	0.656	0.547	N/A	0.305
Gated (Hybrid)	0.609	0.593	0.579	0.621	N/A	0.389
Gated (Spectral)	0.576	0.614	0.532	0.621	N/A	0.390
Stacked (Hybrid)	0.567	0.641	0.665	0.746	N/A	0.495
Stacked (Spectral)	0.558	0.635	0.661	0.744	N/A	0.515
Stacked (Von Neumann)	0.579	0.645	0.651	0.747	N/A	0.523

Table 10: Complete pooled out-of-fold AUROC for GPT-5.1. Missing generation, judging, or modality combinations are shown as N/A; Cheque Generation is reported separately because these agents are text-only.

Method	AA Finance	AmbigQA	HotpotQA	SQuAD	Long Text	Summ. Finance
Standard SE	0.523	0.706	0.579	0.562	0.240	0.500
SE UEigV	0.524	0.705	0.579	0.559	0.176	0.501
SE Hybrid	0.522	0.706	0.578	0.560	0.216	0.500
SE Von Neumann	0.533	0.725	0.661	0.635	0.328	0.496
TopK	0.767	0.682	0.706	0.583	0.712	0.440
Spectral Epistemic (P3)	0.530	0.713	0.579	0.557	0.256	0.500
CoCoA SP	0.608	0.740	0.706	0.681	0.784	0.521
CoCoA PPL	0.615	0.714	0.686	0.661	0.752	0.518
Gated (Hybrid)	0.653	0.696	0.692	0.681	0.216	0.481
Gated (Spectral)	0.659	0.707	0.685	0.706	0.584	0.486
Stacked (Hybrid)	0.799	0.665	0.685	0.735	0.664	0.483
Stacked (Spectral)	0.801	0.682	0.689	0.733	0.624	0.472
Stacked (Von Neumann)	0.804	0.677	0.688	0.733	0.616	0.491

E.2 Response-count ablation

Figure 11 examines sensitivity to the number of sampled responses. Increasing N can expose additional semantic variation and stabilise multi-response token statistics, but the resulting AUROC changes are not uniformly positive across datasets or method categories. The absence of a consistent monotonic improvement indicates that additional samples alone do not resolve the dataset-dependent weaknesses of either signal family.

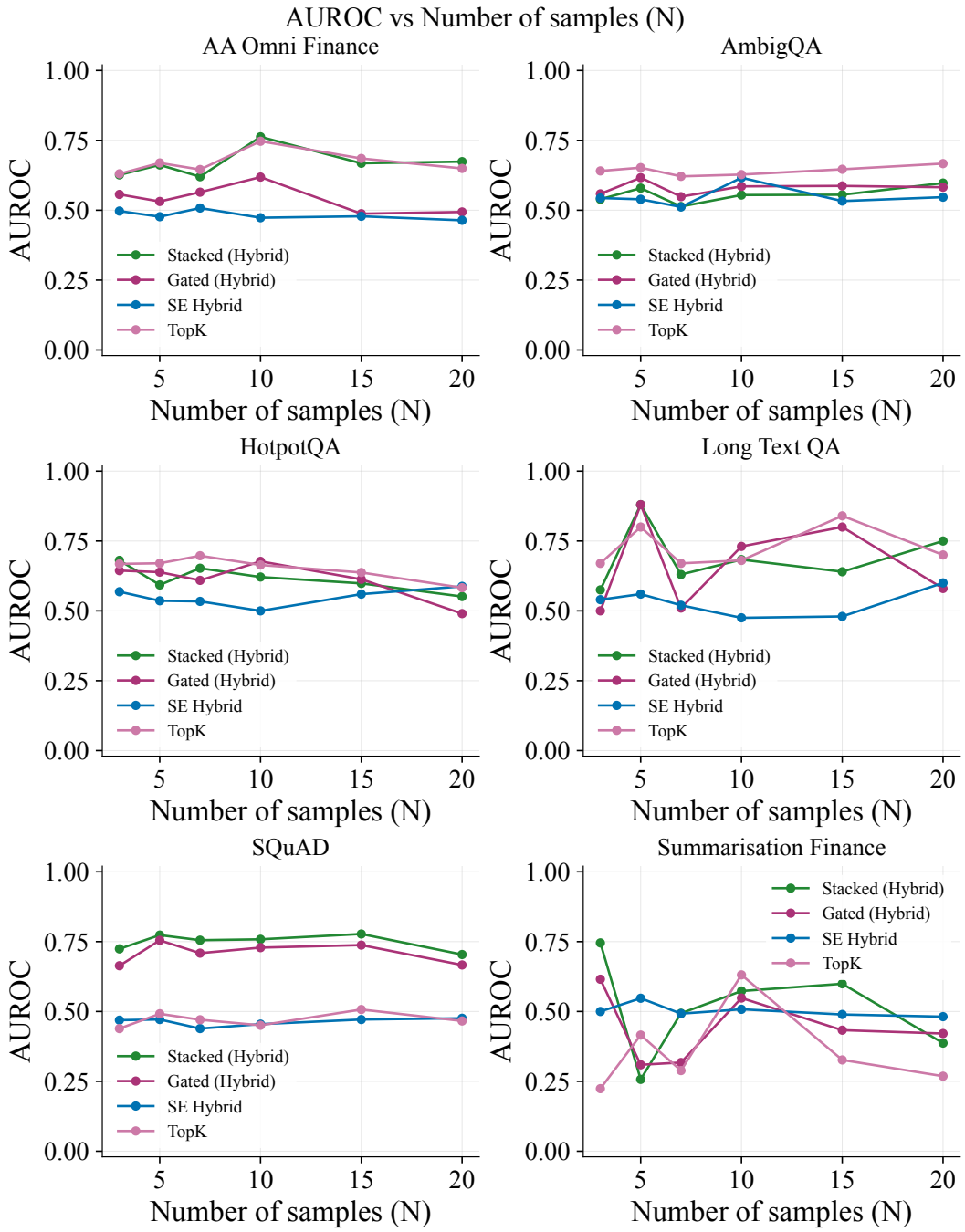

Figure 11: **AUROC versus response count.** Performance is shown over $N \in \{3, 5, 7, 10, 15, 20\}$ for the strongest semantic method, TopK, and the averaged Gated and Stacked variants. The effect of increasing N varies across datasets and methods. CoCoA is omitted because the required main-answer embeddings are unavailable for this sweep.

Table 11: Complete pooled out-of-fold AUROC for GPT-5.4. Missing generation, judging, or modality combinations are shown as N/A; Cheque Generation is reported separately because these agents are text-only.

Method	AA Finance	AmbigQA	HotpotQA	SQuAD	Long Text	Summ. Finance
Standard SE	0.503	0.628	0.624	0.481	0.619	0.492
SE UEigV	0.506	0.636	0.624	0.481	0.595	0.492
SE Hybrid	0.509	0.641	0.628	0.481	0.595	0.492
SE Von Neumann	0.529	0.640	0.666	0.556	0.660	0.722
TopK	0.725	0.633	0.703	0.736	0.585	0.562
Spectral Epistemic (P3)	0.501	0.621	0.629	0.481	0.619	0.492
CoCoA SP	0.584	0.676	0.680	0.623	0.707	0.679
CoCoA PPL	0.593	0.655	0.698	0.613	0.626	0.641
Gated (Hybrid)	0.611	0.651	0.684	0.605	0.469	0.388
Gated (Spectral)	0.587	0.641	0.641	0.606	0.490	0.388
Stacked (Hybrid)	0.698	0.601	0.702	0.588	0.714	0.496
Stacked (Spectral)	0.698	0.627	0.713	0.573	0.694	0.579
Stacked (Von Neumann)	0.694	0.610	0.695	0.574	0.701	0.522

Table 12: Cheque Generation is a multimodal vision-language task. Text-only model-family agents are structurally unavailable rather than failed or imputed. The primary GPT-4.1-mini result and the externally scored N-sweep are reported separately.

Evidence	Coverage	Status
Primary evaluation	GPT-4.1-mini, all detector methods	pooled OOF
N-sweep	GPT-4.1-mini, $N = 3$–20	external evaluator
Model-family sweep	text-only agents	N/A by modality

E.3 Sampling-temperature ablation

Sampling temperature changes both response diversity and token confidence. Figure 12 shows that higher temperatures generally reduce the single-cluster rate, but do not produce a consistent AUROC improvement. Greater diversity can expose alternative meanings, but it can also introduce variation unrelated to hallucination; the preferred temperature is therefore dataset-dependent.

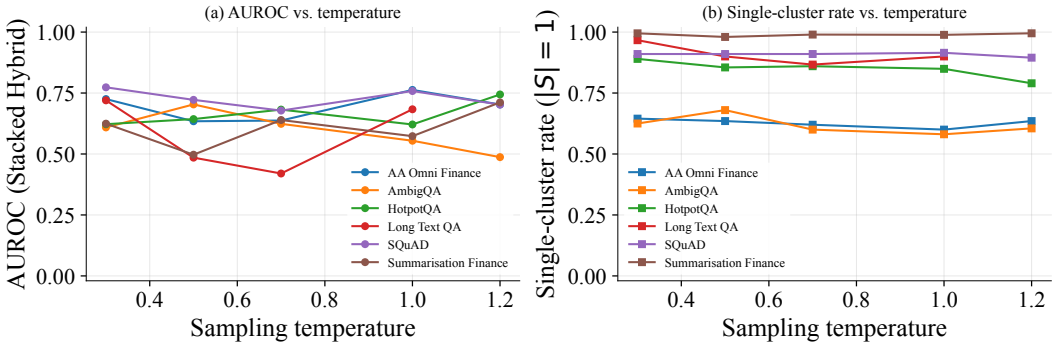

Figure 12: **Sensitivity to sampling temperature.** (a) Detector AUROC and (b) the proportion of queries forming a single semantic cluster over $T \in \{0.3, 0.5, 0.7, 1.0, 1.2\}$. Higher response diversity does not consistently improve hallucination separation. CoCoA is omitted because the required main-answer embeddings are unavailable for this sweep.

E.4 Top-k candidate ablation

Table 13 evaluates the number of token candidates retained at each generation step. Most improvements occur when moving from $k = 1$ to $k = 2$ or $k = 3$; additional candidates produce smaller and less consistent changes. Semantic-only performance is unchanged because it does not use token alternatives, and CoCoA SP/PPL are likewise flat across k: both depend only on the generated answer's own token log-probabilities, never the truncated alternative list. The available responses store at most five candidates, so values above $k = 5$ cannot be reconstructed without new generations.

E.5 Calibration sensitivity

We sweep five clustering thresholds $\tau \in \{0.80, 0.85, 0.90, 0.95, 0.98\}$ and three regularisation strengths $C \in \{0.1, 1, 10\}$, giving 15 calibration settings. Figure 13 shows the method-level minimum–maximum AUROC ranges, with colour indicating interval width; Table 14 gives the corresponding method-category summary. Narrow intervals indicate that a method is comparatively insensitive to calibration, whereas wider intervals identify datasets and method families for which the selected configuration materially affects performance. CoCoA SP and CoCoA PPL have zero-width intervals throughout, alongside SE Von Neumann and TopK: CoCoA has neither a clustering threshold nor a regularisation strength to sweep.

F Statistical significance and uncertainty

This section complements the point-estimate results with significance tests and bootstrap confidence intervals. The significance tests assess whether method scores differ between hallucinated and non-hallucinated queries, while the bootstrap intervals quantify uncertainty in the resulting AUROC estimates.

F.1 Significance tests

Table 15 reports Mann–Whitney U tests on the pooled out-of-fold method scores. All three Stacked variants significantly separate the labels on AA Omni Finance, Cheque Generation, and SQuAD, but not on AmbigQA, HotpotQA, Long-Text QA, or Financial Summaries at $p < 0.05$. Gated Hybrid is significant on five datasets, with Long-Text QA and Financial Summaries as the exceptions.

Table 13: **Top-k log-probability sensitivity, by detector category.** AUROC vs. the number of top-k log-probability alternatives requested per token, recomputed from existing saved responses (no new generation; $k > 5$ not recoverable from already-stored data). Best SE is the strongest of the 4 SE variants per cell; Avg Stacked/Gated average their variants. CoCoA SP/PPL are flat across k because both depend only on the generated answer's own token log-probabilities, never the truncated alternatives.

		AUROC at $k =$				
Dataset	Category	1	2	3	4	5
AA Omni Finance	Best SE	0.543	0.543	0.543	0.543	0.543
AA Omni Finance	TopK	0.621	0.704	0.721	0.728	0.735
AA Omni Finance	Avg Stacked	0.634	0.670	0.741	0.740	0.742
AA Omni Finance	Avg Gated	0.542	0.566	0.587	0.601	0.611
AA Omni Finance	CoCoA SP	0.687	0.687	0.687	0.687	0.687
AA Omni Finance	CoCoA PPL	0.704	0.704	0.704	0.704	0.704
AmbigQA	Best SE	0.652	0.652	0.652	0.652	0.652
AmbigQA	TopK	0.614	0.636	0.638	0.638	0.639
AmbigQA	Avg Stacked	0.554	0.526	0.548	0.546	0.543
AmbigQA	Avg Gated	0.548	0.589	0.586	0.581	0.581
AmbigQA	CoCoA SP	0.667	0.667	0.667	0.667	0.667
AmbigQA	CoCoA PPL	0.654	0.654	0.654	0.654	0.654
Summarisation Finance	Best SE	0.508	0.508	0.508	0.508	0.508
Summarisation Finance	TopK	0.530	0.609	0.615	0.617	0.619
Summarisation Finance	Avg Stacked	0.597	0.538	0.518	0.527	0.556
Summarisation Finance	Avg Gated	0.498	0.538	0.588	0.525	0.540
Summarisation Finance	CoCoA SP	0.513	0.513	0.513	0.513	0.513
Summarisation Finance	CoCoA PPL	0.516	0.516	0.516	0.516	0.516
HotpotQA	Best SE	0.627	0.627	0.627	0.627	0.627
HotpotQA	TopK	0.596	0.655	0.654	0.656	0.656
HotpotQA	Avg Stacked	0.621	0.597	0.614	0.594	0.599
HotpotQA	Avg Gated	0.493	0.639	0.639	0.638	0.640
HotpotQA	CoCoA SP	0.681	0.681	0.681	0.681	0.681
HotpotQA	CoCoA PPL	0.696	0.696	0.696	0.696	0.696
SQuAD	Best SE	0.537	0.537	0.537	0.537	0.537
SQuAD	TopK	0.463	0.452	0.459	0.461	0.460
SQuAD	Avg Stacked	0.767	0.768	0.768	0.760	0.759
SQuAD	Avg Gated	0.535	0.738	0.738	0.739	0.740
SQuAD	CoCoA SP	0.575	0.575	0.575	0.575	0.575
SQuAD	CoCoA PPL	0.546	0.546	0.546	0.546	0.546
Long Text QA	Best SE	0.647	0.647	0.647	0.647	0.647
Long Text QA	TopK	0.696	0.705	0.710	0.705	0.705
Long Text QA	Avg Stacked	0.641	0.608	0.586	0.521	0.611
Long Text QA	Avg Gated	0.525	0.737	0.719	0.701	0.681
Long Text QA	CoCoA SP	0.683	0.683	0.683	0.683	0.683
Long Text QA	CoCoA PPL	0.661	0.661	0.661	0.661	0.661
Cheque Generation	Best SE	0.563	0.563	0.563	0.563	0.563
Cheque Generation	TopK	0.568	0.602	0.598	0.594	0.593
Cheque Generation	Avg Stacked	0.725	0.742	0.741	0.739	0.731
Cheque Generation	Avg Gated	0.451	0.626	0.624	0.622	0.617
Cheque Generation	CoCoA SP	0.618	0.618	0.618	0.618	0.618
Cheque Generation	CoCoA PPL	0.609	0.609	0.609	0.609	0.609

AmbigQA shows the clearest significance for semantic-only methods. CoCoA SP and CoCoA PPL are significant on AA Omni Finance, AmbigQA, Cheque Generation, and HotpotQA; CoCoA SP is additionally significant on Long-Text QA. Neither CoCoA variant reaches significance on SQuAD or Financial Summaries. On Financial Summaries this mirrors the weak, non-significant separation semantic-only methods also show there. On SQuAD, however, four of the five semantic-only methods do reach significance (Standard SE, UEigV, Hybrid, Spectral Epistemic; all $p \leq 0.030$) despite sub-chance AUROC ($\approx$0.48) — indicating a significant but inverted relationship, consistent with the positional-entropy inversion for SQuAD discussed in Section 4.2, rather than an absence of signal.

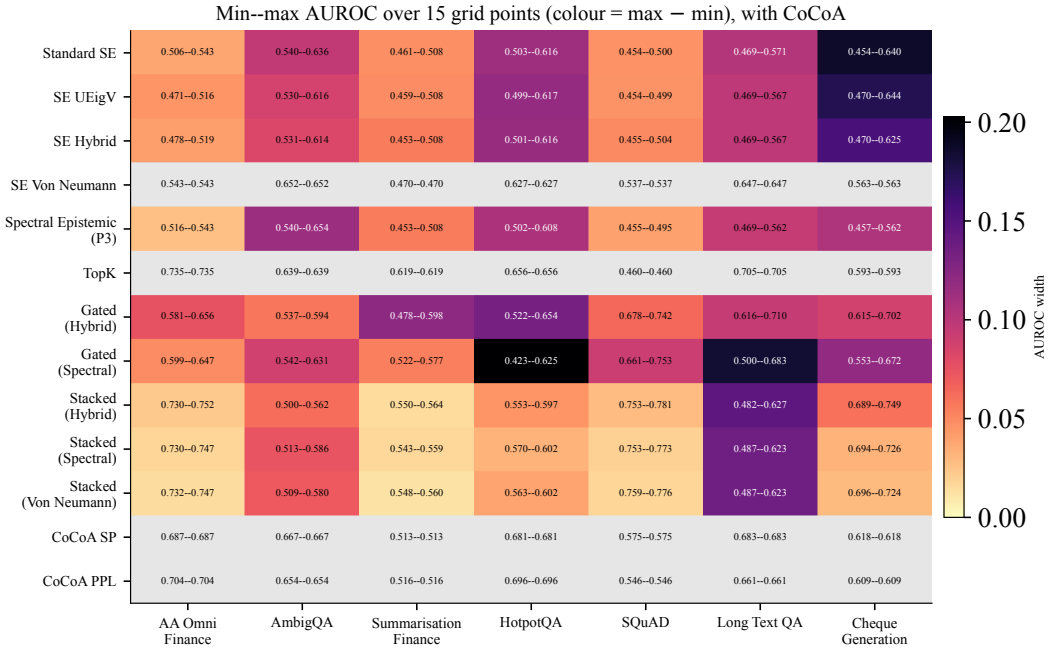

Figure 13: **Calibration sensitivity, with CoCoA.** Cells report the minimum–maximum AUROC over 15 (τ, C) settings; colour encodes the width of that range. CoCoA SP/PPL are included as zero-width reference rows.

These unadjusted tests are exploratory and should be interpreted alongside effect sizes and confidence intervals rather than as evidence that one method is universally superior.

F.2 Bootstrap confidence intervals and AUROC aggregation

Table 7 reports the mean of five held-out fold AUROCs, whereas Table 16 pools the same out-of-fold predictions before computing AUROC. For the pooled estimates, we obtain 95% confidence intervals from 2,000 query-level percentile-bootstrap resamples. The two aggregation procedures produce similar values in most cells, although small differences can change the leading method when methods perform closely.

Aggregation-sensitive rankings. The leading method category is stable under both aggregation procedures on six of seven datasets. Long-Text QA is the exception: mean fold AUROC ranks Gated Hybrid first at 0.731, whereas pooled AUROC ranks CoCoA SP first at 0.728 and Gated Hybrid at 0.688. AmbigQA and SQuAD also exhibit small reversals between variants within the same method category, but the corresponding differences are approximately 0.001. These changes show that close rankings should not be interpreted without their uncertainty intervals.

Interpreting the intervals. The leading estimate overlaps at least one competing method on every dataset. For example, Stacked Hybrid reaches 0.746 on AA Omni Finance with a 95% interval of $[0.67, 0.82]$, overlapping TopK at 0.735 $[0.66, 0.81]$. Uncertainty is especially large for Long-Text QA because it contains only 30 evaluated queries. The results therefore support broad cross-dataset patterns, but do not establish statistically distinct superiority for an individual method on any single benchmark.

Table 14: **Calibration-threshold sensitivity, by detector category.** AUROC sensitivity to clustering threshold τ and regularisation C, swept over $\tau \in \{0.80, \ldots, 0.98\}$, $C \in \{0.1, 1, 10\}$, for each detector category (Best SE = strongest of the 4 SE variants at that grid point; Avg Stacked/Gated average their variants). The reported interval is the minimum–maximum AUROC over all 15 grid points. CoCoA SP/PPL have zero-width ranges because CoCoA has neither a clustering threshold nor a regularisation strength to sweep.

Dataset	Category	AUROC range over 15 settings
AA Omni Finance	Best SE	0.543–0.543
AA Omni Finance	TopK	0.735–0.735
AA Omni Finance	Avg Stacked	0.731–0.748
AA Omni Finance	Avg Gated	0.590–0.648
AA Omni Finance	CoCoA SP	0.687–0.687
AA Omni Finance	CoCoA PPL	0.704–0.704
AmbigQA	Best SE	0.652–0.652
AmbigQA	TopK	0.639–0.639
AmbigQA	Avg Stacked	0.509–0.576
AmbigQA	Avg Gated	0.540–0.610
AmbigQA	CoCoA SP	0.667–0.667
AmbigQA	CoCoA PPL	0.654–0.654
Summarisation Finance	Best SE	0.470–0.508
Summarisation Finance	TopK	0.619–0.619
Summarisation Finance	Avg Stacked	0.549–0.561
Summarisation Finance	Avg Gated	0.520–0.588
Summarisation Finance	CoCoA SP	0.513–0.513
Summarisation Finance	CoCoA PPL	0.516–0.516
HotpotQA	Best SE	0.627–0.627
HotpotQA	TopK	0.656–0.656
HotpotQA	Avg Stacked	0.566–0.599
HotpotQA	Avg Gated	0.472–0.640
HotpotQA	CoCoA SP	0.681–0.681
HotpotQA	CoCoA PPL	0.696–0.696
SQuAD	Best SE	0.537–0.537
SQuAD	TopK	0.460–0.460
SQuAD	Avg Stacked	0.755–0.777
SQuAD	Avg Gated	0.675–0.747
SQuAD	CoCoA SP	0.575–0.575
SQuAD	CoCoA PPL	0.546–0.546
Long Text QA	Best SE	0.647–0.647
Long Text QA	TopK	0.705–0.705
Long Text QA	Avg Stacked	0.485–0.624
Long Text QA	Avg Gated	0.574–0.683
Long Text QA	CoCoA SP	0.683–0.683
Long Text QA	CoCoA PPL	0.661–0.661
Cheque Generation	Best SE	0.563–0.644
Cheque Generation	TopK	0.593–0.593
Cheque Generation	Avg Stacked	0.693–0.731
Cheque Generation	Avg Gated	0.590–0.683
Cheque Generation	CoCoA SP	0.618–0.618
Cheque Generation	CoCoA PPL	0.609–0.609

Table 15: Mann–Whitney U test p-values for separation between hallucinated and non-hallucinated out-of-fold scores. The available artifact does not contain a Gated Spectral significance row. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Method	AA Fin.	AmbigQA	Cheque	HotpotQA	Long QA	SQuAD	Summ. Fin.
Standard SE	0.900	0.002**	0.691	0.934	0.604	0.028*	0.547
SE UEigV	0.487	0.009**	0.833	0.989	0.604	0.027*	0.547
SE Hybrid	0.608	0.011*	0.988	0.985	0.604	0.029*	0.547
SE Von Neumann	0.607	0.001**	0.254	0.021*	0.329	0.233	0.588
TopK	$< 10^{-4}$***	0.004**	0.091	0.004**	0.059	0.354	0.027*
Spectral Epistemic	0.705	$< 10^{-3}$***	0.625	0.952	0.661	0.030*	0.547
CoCoA SP	0.0002***	0.0003***	0.032*	0.0009***	0.036*	0.115	0.821
CoCoA PPL	$< 10^{-4}$***	0.0003***	0.045*	0.0005***	0.059	0.347	0.692
Gated Hybrid	0.038*	0.029*	0.021*	0.004**	0.084	$< 10^{-4}$***	0.125
Gated Spectral	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Stacked Hybrid	$< 10^{-4}$***	0.487	$< 10^{-4}$***	0.071	0.318	$< 10^{-4}$***	0.435
Stacked Spectral	$< 10^{-4}$***	0.215	$< 10^{-4}$***	0.057	0.280	$< 10^{-4}$***	0.484
Stacked Von Neumann	$< 10^{-4}$***	0.230	$< 10^{-3}$***	0.076	0.299	$< 10^{-4}$***	0.335

Table 16: **Pooled out-of-fold AUROC with 95% bootstrap confidence intervals** (2000 resamples; percentile method) for every (dataset, method) cell in Table 7. Point estimates here are the pooled-OOF AUROC, which can differ slightly from Table 7's mean-of-fold AUROC (see that table's footnote and §F.2 for the three datasets where this changes the outright winner).

Method	AA Omni Finance	AmbigQA	Financial Summaries	HotpotQA	SQuAD	Long-Text QA	Cheque Generation
Standard SE	0.506 [0.42, 0.59]	0.632 [0.54, 0.72]	0.504 [0.50, 0.51]	0.503 [0.43, 0.57]	0.454 [0.41, 0.50]	0.469 [0.36, 0.58]	0.483 [0.40, 0.56]
SE UEigV	0.471 [0.39, 0.55]	0.612 [0.52, 0.69]	0.504 [0.50, 0.51]	0.499 [0.43, 0.56]	0.454 [0.41, 0.50]	0.469 [0.36, 0.58]	0.492 [0.42, 0.56]
SE Hybrid	0.478 [0.39, 0.57]	0.610 [0.52, 0.69]	0.504 [0.50, 0.51]	0.501 [0.43, 0.56]	0.455 [0.41, 0.50]	0.469 [0.36, 0.58]	0.499 [0.42, 0.57]
SE Von Neumann	0.525 [0.43, 0.62]	0.653 [0.56, 0.74]	0.473 [0.38, 0.57]	0.625 [0.51, 0.73]	0.551 [0.46, 0.64]	0.607 [0.39, 0.84]	0.562 [0.46, 0.66]
TopK	0.735 [0.66, 0.81]	0.636 [0.54, 0.73]	**0.610 [0.51, 0.70]**	0.656 [0.55, 0.76]	0.460 [0.37, 0.55]	0.705 [0.50, 0.90]	0.593 [0.49, 0.69]
Spectral Epistemic (P3)	0.516 [0.43, 0.60]	0.652 [0.56, 0.74]	0.504 [0.50, 0.51]	0.502 [0.43, 0.57]	0.455 [0.41, 0.50]	0.473 [0.36, 0.58]	0.479 [0.40, 0.56]
CoCoA SP	0.685 [0.60, 0.77]	0.672 [0.57, 0.76]	0.511 [0.42, 0.60]	0.679 [0.57, 0.78]	0.567 [0.48, 0.65]	**0.728 [0.53, 0.92]**	0.617 [0.51, 0.71]
CoCoA PPL	0.697 [0.62, 0.78]	**0.673 [0.57, 0.76]**	0.520 [0.43, 0.61]	**0.689 [0.59, 0.78]**	0.540 [0.45, 0.63]	0.705 [0.50, 0.90]	0.609 [0.50, 0.71]
Gated (Hybrid)	0.602 [0.51, 0.69]	0.604 [0.51, 0.70]	0.576 [0.48, 0.67]	0.654 [0.56, 0.75]	0.729 [0.65, 0.80]	0.688 [0.47, 0.88]	0.626 [0.54, 0.71]
Gated (Spectral)	0.620 [0.53, 0.70]	0.605 [0.51, 0.70]	0.575 [0.48, 0.67]	0.625 [0.53, 0.73]	0.752 [0.68, 0.82]	0.674 [0.44, 0.88]	0.608 [0.51, 0.69]
Stacked (Hybrid)	**0.746 [0.67, 0.82]**	0.533 [0.44, 0.63]	0.539 [0.44, 0.63]	0.597 [0.48, 0.71]	0.759 [0.68, 0.83]	0.609 [0.38, 0.82]	**0.749 [0.65, 0.84]**
Stacked (Spectral)	0.742 [0.66, 0.81]	0.559 [0.46, 0.65]	0.535 [0.44, 0.63]	0.603 [0.49, 0.72]	0.759 [0.68, 0.83]	0.618 [0.39, 0.83]	0.725 [0.62, 0.81]
Stacked (Von Neumann)	0.739 [0.66, 0.81]	0.557 [0.46, 0.65]	0.548 [0.45, 0.64]	0.596 [0.48, 0.71]	**0.760 [0.68, 0.83]**	0.614 [0.38, 0.82]	0.720 [0.62, 0.81]