How Do Uncertainty Estimators Align with Hallucination Occurrence in LLMs?

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) can hallucinate, but do model uncertainty estimators reliably reveal when they do? Although uncertainty estimators are increasingly used to identify unreliable generations, their relationship with hallucination remains poorly understood and is often taken for granted. We systematically evaluate 46 uncertainty estimators across complementary hallucination settings, covering both intrinsic hallucinations, where outputs are unsupported by the provided context, and extrinsic hallucinations, where claims are unsupported by expected model knowledge. Across four benchmarks and multiple LLMs, we find that the association between uncertainty and hallucination varies widely and is often weak, depending on the hallucination setting and the model under evaluation. These results challenge the use of uncertainty as a direct signal of hallucination and clarify when it provides useful information.

1 Introduction

Large language models (LLMs) have demonstrated strong performance across a wide range of natural language generation tasks (Zhao et al., 2025). However, fluent and coherent responses are not necessarily supported by the evidence the model is expected to rely on. Such unsupported statements, commonly referred to as hallucinations, remain an important obstacle to the reliable deployment of LLMs (Bang et al., 2025; Fan et al., 2026), particularly in settings where errors can have serious consequences (Bengio et al., 2025; Sahoo et al., 2024).

Assessing hallucination first requires specifying the relevant source of evidence. Recent work distinguishes hallucination from factuality, which concerns correctness relative to real-world truth (Bang et al., 2025; Augenstein et al., 2024). Following Bang et al. (2025), we distinguish intrinsic hallucinations, where model responses are unsupported by information supplied at inference time, from extrinsic hallucinations, where responses are unsupported by knowledge the model is expected to have encountered during pretraining. These notions are related but not equivalent: a statement can be factually correct while remaining unsupported by the context it was expected to use. This distinction matters when evaluating methods intended to identify hallucinated responses.

Substantial effort has been devoted to uncertainty estimation (UE) for language models, with the aim of quantifying confidence in generated responses (Vashurin et al., 2025a; Kang et al., 2025). Existing methods rely on different sources of information, including token probabilities and likelihoods (Fomicheva et al., 2020; Malinin & Gales, 2021), variability across sampled generations (Kuhn et al., 2023; Lin et al., 2023), internal model representations, and model self-evaluation (Kadavath et al., 2022; Tian et al., 2023). In practice, such signals can support selective prediction or abstention, route uncertain responses to additional verification, or trigger retrieval and refinement mechanisms (Abbasi-Yadkori et al., 2024; Yang et al., 2023; Yao et al., 2024; Moskvoretskii et al., 2025).

Although uncertainty estimators are not specifically designed to detect hallucinations, they are widely used for this purpose. Yet, the extent to which their scores reflect hallucination remains insufficiently understood. Most prior evaluations assess uncertainty against broader notions of correctness, generation quality, robustness, or calibration rather than against explicit hallucination criteria (Vashurin et al., 2025a; Bakman et al., 2025; Wang et al., 2025). Moreover, uncertainty estimators capture different signals, and their ability to detect one type of failure does not necessarily extend to other forms of hallucination. Recent work has also shown that estimator comparisons are sensitive to evaluation choices, including correctness functions, and to confounders such as response

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

length (Santilli et al., 2025; Ielanskyi et al., 2025). Consequently, existing results provide limited insights into when and under what conditions uncertainty scores reliably indicate hallucination.

In this paper, we directly evaluate the relationship between uncertainty estimators and hallucination. We compare 46 estimators across 4 datasets (tasks) and 3 open-weight instruction-tuned models. The tasks cover context faithfulness in retrieval-augmented generation (intrinsic) and extrinsic hallucinations, encompassing short-form question answering, long-form generation, and abstention on false premises. This design allows us to examine not only whether uncertainty is associated with hallucination, but also how this association changes across task types and models.

Our study makes three main contributions. First, we systematically evaluate a broad set of uncertainty estimators for hallucination detection across intrinsic and extrinsic hallucination settings. Second, we characterize the relationship between uncertainty and hallucination across tasks and models. We find substantial variation in this relationship: uncertainty provides a strong signal across multiple estimators in some settings, whereas in others, it is useful only for a smaller subset of estimators or varies considerably across models. Third, we investigate estimator selection beyond individual performance by examining the stability of estimator rankings across tasks and models and the relationships between competitive uncertainty scores. We find that estimator rankings vary substantially across settings, although several poorly performing estimators consistently rank near the bottom. Moreover, some competitive estimators produce strongly correlated response rankings despite differing computational requirements. Together, these findings show that the utility of uncertainty for hallucination detection depends on the evaluation setting and highlight the importance of evaluating estimators under the specific conditions in which they are intended to be used.

2 Background on Uncertainty Estimators

Notation. We consider an autoregressive language model $p_{\boldsymbol{\theta}}(y \mid x)$, where x denotes the input (query and optional context), $y = (y_1, \ldots, y_T)$ is a generated response sequence, and $\boldsymbol{\theta}$ parametrizes the model. At each step t, the model produces a token-level probability distribution $p_{\boldsymbol{\theta}}(\cdot \mid x, y_{<t})$, so that $p_{\boldsymbol{\theta}}(y \mid x) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(y_t \mid x, y_{<t})$ for a response sequence of length T, with $y_{<t} = (y_1, \ldots, y_{t-1})$.

2.1 Uncertainty Definition and Scope

There is no universally accepted formal definition of uncertainty in language generation (Zhang et al., 2024; Bakman et al., 2025). In this work, uncertainty is treated as a query-level property of the predictive distribution $p_{\boldsymbol{\theta}}(\cdot \mid x)$. A generic query-level target is denoted $U^{\star}(x) = \Phi(p_{\boldsymbol{\theta}}(\cdot \mid x))$, where Φ is a property of the predictive distribution (Chen et al., 2026). Different estimators may approximate different aspects of this target and need not share the same probabilistic foundation.

Terminology in the literature is not fully consistent: some works use *uncertainty* and *confidence* interchangeably (Qiu & Miikkulainen, 2024; Zhang et al., 2024), while others reserve *confidence* for response-level uncertainty y (Lin et al., 2023). Here, *uncertainty* refers by default to query-level uncertainty. Scores of the form $U(x)$ target this quantity directly, whereas response-conditioned scores $U(x, y)$ are treated as single-response proxies. Our empirical evaluation associates each score with the quality of the observed response; it does not estimate $U^{\star}(x)$ independently.

2.2 Uncertainty Estimators Studied

We organize uncertainty estimators by their access requirements: white-box [WB] methods require access to the evaluated model's logits or internal states, whereas black-box [BB] methods operate on generated text alone. Within each access regime, we group estimators by the information used to compute uncertainty and their underlying methodology. Our implementation follows LM-Polygraph (Fadeeva et al., 2023; Vashurin et al., 2025a); see detailed definitions in Section B.

[WB] Information/logit-based methods estimate uncertainty from token-level predictive distributions $p_{\boldsymbol{\theta}}(\cdot \mid x, y_{<t})$, without requiring additional sampled responses. They differ in what information they extract from these distributions and how they aggregate it across tokens. The simplest methods, such as Maximum Sequence Probability (MSP), use only the probabilities assigned to generated

tokens (Fomicheva et al., 2020). Other methods exploit the full token distribution, measuring uncertainty through token entropy (Fomicheva et al., 2020) or measures of deviation from the uniform distribution, such as Rényi or Fisher–Rao divergences (Darrin et al., 2023). Pointwise mutual information methods compare conditional and unconditional token probabilities to quantify the influence of the input on the generated response (van der Poel et al., 2022). Finally, some methods incorporate token importance: TokenSAR weights token-level uncertainty by semantic importance (Duan et al., 2024), while Claim-Conditioned Probability (CCP) assesses uncertainty based on whether alternative tokens preserve or alter the meaning of the current claim (Fadeeva et al., 2024).

[WB] Sample-based methods require multiple responses sampled from the evaluated model, together with their token-level probabilities. They estimate uncertainty from the variability of sampled responses, using either likelihood-based or semantic information. Monte Carlo methods approximate predictive uncertainty through sample-based estimates of entropy or related quantities (Malinin & Gales, 2021). Semantic Entropy (Kuhn et al., 2023) groups responses into semantic equivalence classes and measures uncertainty over these classes, whereas Semantic Density (Qiu & Miikkulainen, 2024) uses continuous pairwise semantic similarities. SentenceSAR (Duan et al., 2024) combines sequence likelihoods with semantic similarities to weight sampled responses, while Cocoa (Vashurin et al., 2025b) estimators combine logit-based uncertainty with semantic variability across samples.

[WB] Internal-state methods require access to the model's attention weights or hidden representations. They estimate uncertainty from internal processing signals rather than solely from output probabilities. AttentionScore uses self-attention patterns (Sriramanan et al., 2024), RAUQ models uncertainty propagation across the generated sequence (Vazhentsev et al., 2025), and attention-based saliency methods reweight token probabilities according to their contextual importance (Lin et al., 2024a).

[WB] Training-based methods require internal representations and a separate dataset to fit a reference distribution. They estimate uncertainty by measuring how far an input or response representation lies from the fitted distribution, typically using Gaussian density models and Mahalanobis distance (Vazhentsev et al., 2023; Lee et al., 2018). Unlike the preceding methods, they require a fitting stage before evaluation.

[WB] Reflexive methods estimate uncertainty from the model's explicit assessment of its own responses, typically elicited through a follow-up prompt. For example, $P(\mathrm{True})$ uses the probability assigned to a positive answer when the model is asked to assess the correctness of its generated response (Kadavath et al., 2022). Such methods rely on the model's self-assessment rather than directly aggregating the original response's predictive probabilities.

[BB] Black-box methods require only generated text, without access to the evaluated model's logits or internal states. Most rely on multiple sampled responses and estimate uncertainty from their lexical or semantic agreement (Lin et al., 2023; Nikitin et al., 2024). Pairwise similarities, computed using lexical overlap or semantic similarity measures such as natural language inference, are aggregated into uncertainty scores through clustering, graph-based statistics, or spectral analysis. Unlike white-box methods, they can be applied without access to the evaluated model's internal computations.

3 Benchmark Setup

3.1 Datasets

Our evaluation covers both intrinsic hallucinations, assessed against evidence provided with the input, and extrinsic hallucinations, assessed against knowledge expected to be available from pretraining. We consider four tasks targeting different aspects of hallucination: context faithfulness, short-answer correctness, factual support in long-form generation, and abstention on false premises. Each task defines a response-level quality target $Q(x, y)$ based on its specific evaluation criterion. We use these targets to assess whether higher uncertainty is associated with lower response quality.

Intrinsic hallucination: RAGTruth . RAGTruth (Niu et al., 2024) evaluates context faithfulness across summarization, passage-grounded question answering, and data-to-text generation. Human annotators identify response spans that conflict with or are unsupported by the provided context. We

use the processed release[1] and define $Q(x, y) = 0$ if the response contains at least one annotated hallucinated span, and $Q(x, y) = 1$ otherwise. Under this criterion, a statement may be considered hallucinated even if it is factually correct, provided that it is not supported by the supplied context.

Extrinsic hallucination: HalluLens . We consider three tasks from HalluLens (Bang et al., 2025), which use Wikipedia as a proxy for knowledge expected to be available from pretraining and synthetic nonexistent entities to evaluate abstention when the requested information does not exist. We follow the HalluLens evaluation procedures, using Llama-3.1-70B-Instruct[2] as the judge. Label processing, quality formulas and evaluated response counts are provided in Section C.1.

PreciseWikiQA evaluates short-answer correctness in closed-book question answering. Questions and reference answers are generated from sampled Wikipedia pages using an LLM. The judge assesses the correctness of each response against its reference answer. We retain responses with a valid correctness judgment and no abstention flag, assigning $Q(x, y) = 1$ to correct answers and $Q(x, y) = 0$ to incorrect answers. **LongWiki** evaluates the factual support and quantity of information in long-form responses. An LLM decomposes each response into atomic claims, which are then evaluated against retrieved Wikipedia passages using the judge. We measure response quality using F1@K, the harmonic mean of claim-level precision and recall@K. Precision measures the proportion of generated claims that are supported by the reference evidence, while recall@K measures the number of supported claims relative to a fixed quota of K claims. Following Bang et al. (2025), we use $Q(x, y) = \text{F1@}K$, with $K = 32$ throughout our experiments. Further details are provided in Section C.1. **NonExistentRefusal** evaluates whether a model abstains when asked about a nonexistent entity. The judge assesses whether the response rejects the false premise, assigning $Q(x, y) = 1$ when the model abstains and $Q(x, y) = 0$ when it accepts the premise. The quality target therefore measures failure to abstain rather than directly identifying fabricated statements.

3.2 Choice of models and computation of uncertainty estimators

Models. The model selection is constrained by RAGTruth , which provides human-annotated responses generated by six language models (Niu et al., 2024). We retain the three small open-weight models: `Mistral-7B-Instruct-v0.2`, `Llama-2-7b-chat-hf`, and `Llama-2-13b-chat-hf`. Their open weights provide access to token-level probabilities, enabling the evaluation of white-box uncertainty estimators under a common protocol.

Recovering token logits for annotated RAGTruth responses. Because RAGTruth provides responses generated during dataset construction and subsequently annotated for hallucinations, we must evaluate uncertainty estimators on these original responses. Regenerating responses would break their correspondence with the human annotations. However, the dataset does not provide token logits, which are required to compute information-based and certain white-box sampling-based estimators. We therefore recover the necessary token-level probabilities by scoring each annotated response under the corresponding causal language model using teacher forcing (Williams & Zipser, 1989). Given an input–response pair (x, y), we run the model on the concatenated sequence $x \oplus y$ and extract the logits at the response positions. This yields the autoregressive conditional distributions $p_{\boldsymbol{\theta}}(\cdot \mid x, y_{<t})$, including the probabilities $p_{\boldsymbol{\theta}}(y_t \mid x, y_{<t})$ assigned to the observed tokens, which are used to compute the uncertainty estimators on the annotated responses.

Dataset splits. Following Vashurin et al. (2025a), RAGTruth is partitioned into 1000 training and 2000 evaluation queries, keeping all responses associated with a query in the same split. Each HalluLens dataset has 1000 training and 2000 evaluation examples. The training split is used only by estimators requiring fitting or training-set statistics. Summary statistics are reported in Table 1.

Response-length baseline. We include generated token count as a baseline to assess whether uncertainty estimators provided additional predictive information beyond response length alone.

Auxiliary reference-based score: AlignScore. AlignScore AlignScore (Zha et al., 2023) is a trained text-pair model that estimates whether one text is supported by another. We use it as a comparator to contextualize uncertainty estimator performance against a method with access to external reference information. Reference choices and scoring directions are detailed in Section C.2.

[1] huggingface.co/datasets/wandb/RAGTruth-processed

[2] `hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4`.

Table 1: Summary statistics of the benchmark datasets. Token counts with `Llama-2`'s tokenizer. *For RAGTRUTH, the reported target length corresponds to the average generated response length. NONEXISTENTREFUSAL has no reference target answer because the desired behavior is abstention.

Dataset	Train + eval queries	Avg. document length	Avg. target length
PRECISEWIKIQA	$1000 + 2000$	37.9	5.8
LONGWIKI	$1000 + 2000$	38.4	235.0
NONEXISTENTREFUSAL	$1000 + 2000$	17.6	–
RAGTRUTH	$1000 + 1960$	698.2	194.4*

4 EVALUATION METHODOLOGY

Let $\mathcal{D}_{\text{test}} = \{(x_j, y_j)\}_{j=1}^{n}$ denote an evaluation set, where x_j is the input and y_j the model response. Each dataset defines a response-level quality target $Q(x_j, y_j)$, with higher values indicating better responses (Section 3.1). Each uncertainty estimator produces a scalar score u_j for the same instance. We assess whether higher uncertainty is associated with lower quality through discrimination, selective prediction, and rank calibration, then compare estimator rankings and the information their scores capture. For each panel, we estimate metric variability using 1,000 bootstrap resamples and report the resulting standard deviation. For paired AUROC differences, we report 95% percentile confidence intervals using identical resamples for both estimators (Section D.4).

4.1 DETECTION AND SELECTIVE PREDICTION

Discriminating errors from supported responses. The area under the receiver operating characteristic curve (AUROC) measures the probability that an error receives higher uncertainty than a non-error; a random estimator has an expected AUROC of 0.5, and 1 indicates perfect discrimination. We define errors by $Q = 0$ for binary targets and $\text{F1@32} < 0.5$ for LongWiki. Sensitivity to the F1@32 threshold is examined in Section D.7; continuous F1@32 is retained for selective prediction and rank calibration. We also report the area under the precision–recall curve (AUPRC) to complement AUROC when errors and non-errors are imbalanced. As responses are selected from highest to lowest uncertainty, precision is the fraction of selected responses that are errors, and recall is the fraction of all errors selected. A constant score yields AUPRC equal to error prevalence.

Selective prediction. A complementary axis evaluates whether rejecting high-uncertainty responses improves the quality of those retained. Responses are ordered from lowest to highest uncertainty. At rejection rate r, we retain the lowest-uncertainty fraction $1 - r$ and compute its mean quality, defining a quality–retention curve. The oracle instead orders responses by decreasing quality, perfectly rejecting the lowest-quality responses first. Following prior work (Malinin & Gales, 2021; Vashurin et al., 2025a), we summarize the curve with the prediction–rejection ratio (PRR):

$$\text{PRR} = \frac{\text{AUC}_U - \text{AUC}_{\text{random}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{random}}}.$$

Here AUC_U, $\text{AUC}_{\text{random}}$, and $\text{AUC}_{\text{oracle}}$ are the areas under the quality–retention curves for the uncertainty estimator, a random estimator, and the oracle, respectively. All three use rejection rates from 0 to 1. PRR measures the proportion of the oracle's improvement over random rejection that the uncertainty estimator achieves: 1 corresponds to oracle-level performance, 0 to random rejection, and negative values to worse-than-random performance.

Rank calibration. Discrimination and selective prediction assess separability and rejection utility; rank calibration asks whether uncertainty is consistently ordered against expected response quality. Rank-calibration error (RCE; Huang et al., 2024b) provides a rank-based diagnostic of this alignment. It compares scores on different scales without requiring them to be probabilities or binarizing continuous quality targets. Lower RCE is better.

4.2 RANKING STABILITY AND ESTIMATOR REDUNDANCY

After evaluating individual performance, we examine two distinct questions: whether estimator rankings remain stable across evaluation settings, and whether different estimators assign similar rankings to the same responses.

Ranking stability across settings. For each dataset–model pair, referred to as a panel, we rank estimators by their evaluation performance. We then compare these rankings between panels using Kendall's τ, which measures agreement in the ordering of estimator pairs. Holding the model fixed assesses transfer across datasets; holding the dataset fixed assesses stability across models.

Estimator-level correlation. Within each panel, we compute pairwise Spearman correlations between uncertainty scores to measure how similarly estimators rank the same responses. We average these correlations across panels and apply hierarchical clustering to identify groups with similar behavior. This analysis concerns overlap in response rankings; it does not establish whether combining estimators would improve detection.

5 RESULTS

We report results on four datasets covering different hallucination types, RAGTruth (RT), PreciseWikiQA (PWQA), LongWiki (LW), and NonExistentRefusal (NR), and three models: `Mistral-7B-Instruct-v0.2` (M7B), `Llama-2-7B-chat` (L7B), and `Llama-2-13B-chat` (L13B). This yields 12 dataset-model pairs, referred to as *panels*. Unless stated otherwise, estimator rankings exclude AlignScore variants, which are used as reference-based comparators. We use AUROC to compare detection performance, then examine how rankings vary across settings, whether competitive scores are redundant, and how the comparisons depend on targets and controls.

Average performance and consistency across panels. Figure 1 compares each estimator's mean AUROC across panels with the variability of its within-panel rank. Estimators in the upper part of the plot discriminate hallucinations more strongly on average; estimators on the left rank more consistently across panels. CCP (logit/information-based) and CocoaMSP (white-box sample-based) provide the best compromise between the two criteria. MSP (logit/information-based) and Attention-Score (internal-state) achieve comparable or higher mean AUROC but exhibit greater rank variability, indicating that their averages are driven by panels where they peak rather than by uniform performance across panels. Uncertainty is informative on some panels, but no estimator is uniformly reliable. We next compare rank variation across datasets and models.

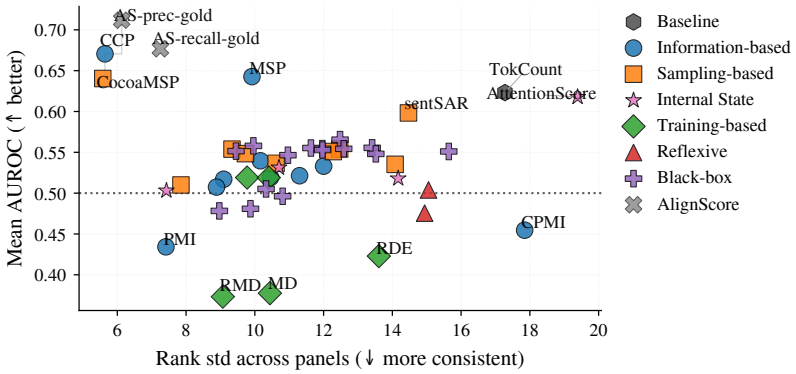


Figure 1: Mean AUROC against rank variability across the 12 panels. Each point represents an uncertainty estimator, the token-count baseline, or an AlignScore variant. Rank variability is the standard deviation of an estimator's within-panel rank across panels; lower values indicate a more stable cross-panel ranking. AlignScore variants are reference-based comparators.

The dataset is the primary source of heterogeneity in estimator rankings. Estimator rankings vary more with dataset changes than with model changes. Kendall's τ heatmaps in Section D.2 (see appendix) quantify this pattern: when the model is fixed and the dataset changes, mean estimator-ranking agreement is low ($\tau = 0.09$ to 0.13); when the dataset is fixed and the model changes, agreement is higher but uneven, with stronger transfer on LW ($\tau = 0.71$) and PWQA ($\tau = 0.66$) than on RT ($\tau = 0.34$) and NR ($\tau = 0.30$).

Family-level comparisons summarize this difference across groups of related estimators. Figure 2 reports family-level predictive performance of uncertainty estimators for two contrasting hallucination

settings: PWQA, where most estimators exhibit moderate to strong discriminative power, and RT, where it is sparsest (see Section 2.2 for the family definitions). On PWQA, all six uncertainty estimator families lie above the 0.5 baseline of an uninformative score, with the information-based and white-box sample-based families reaching the highest family-aggregate AUROCs. In contrast, family-level ROC curves on RT sit closer to the 0.5 baseline, suggesting that uncertainty estimators carry limited information for discriminating hallucinations in this setting. The two remaining datasets, LW and NR, are reported in Figure 7 (see appendix).

Model differences still affect which estimators perform best. Although task differences dominate overall, the model still changes which uncertainty signal is most informative on a given dataset. On RT with M7B, the leading estimators include the information-based MSP and CCP, as well as the white-box, sample-based CocoaMSP (see Table 7 in appendixev). This pattern does not persist on the Llama models. On RT with L13B, by contrast, the output-length baseline is the strongest signal (AUROC 0.75), followed by AttentionScore (0.69) and the reflexive estimators, in which the language model is prompted to assess its own uncertainty (Black-box PTrue 0.67, PTrue 0.63); the reflexive estimators are uninformative on the other models. Thus, even within a fixed task setting, estimator choice remains model-dependent.

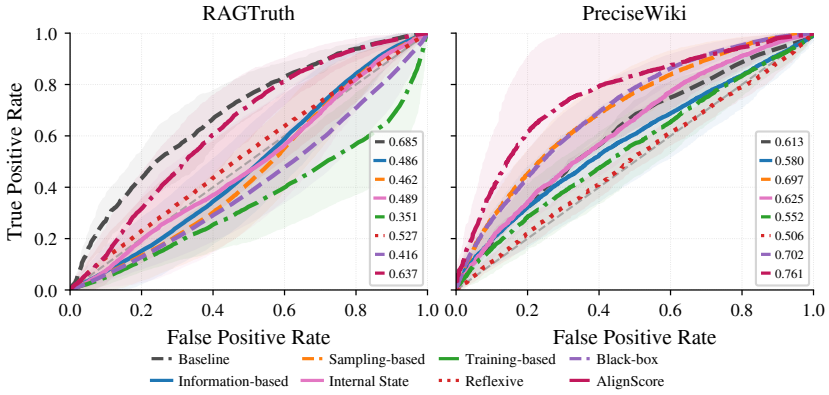


Figure 2: Family-level ROC aggregates per task, averaged across the three generators. For each task–model panel, estimator ROC curves are first averaged within each family by taking the mean true positive rate at each false positive rate. For each task, these model-level family means are then averaged across models. Shaded bands show one standard deviation of these model-level family-mean true positive rates at each false positive rate; they reflect cross-model variability.

Some families consistently perform poorly. The bottom of the panel ranking in Figure 1 is more stable than the top across datasets and models (see Figure 6 in the appendix). The training-based density estimators Mahalanobis Distance (MD), Relative MD, and Robust Density Estimation (RDE) occur among the bottom five uncertainty estimators in 10, 10, and 7 of the 12 panels, respectively, with the pointwise mutual information variants (PMI and CPMI) also frequently near the bottom. A plausible interpretation is that density-based estimators measure how typical a generated response is under a fitted reference representation distribution. Since hallucinated responses are still sampled from the generator, they need not be atypical in representation space. Performance may also depend on modeling choices such as the training data, background corpus, representation layer and dimensionality reduction. These results suggest that training-based density estimators and PMI/CPMI should not be used as default uncertainty signals without dataset-specific validation.

The recurrent top estimators form three correlated ranking clusters. Figure 3 reports Spearman's ρ correlations between the seven estimators with the highest mean AUROC, averaged across panels. This figure shows three correlated ranking clusters. The first cluster is logit- and sample-based, and relies on logits: SentenceSAR, Maximum Sequence Probability (MSP), Claim-Conditioned Probability (CCP), and CocoaMSP. The strongest correlations occur between MSP, CCP, and CocoaMSP ($\rho \geq 0.78$), while SentenceSAR is more weakly linked to the others. This correlation is informative because these estimators differ in construction: MSP is a direct sequence-likelihood score; CCP filters local token alternatives through a natural-language inference model;

CocoaMSP modulates an MSP-derived score using inter-sample semantic dissimilarity; and SentenceSAR combines token-level and sentence-level relevance signals across sampled responses. The second cluster consists of Token count and AttentionScore, which are correlated at $\rho = 0.61$, while their correlations with the probability-based cluster are lower. AttentionScore uses attention weights rather than token probabilities or comparisons between sampled responses. Its correlation with token count may be due to the fact that the scores sum attention token contributions. Correlation alone does not establish that length explains its detection performance or that combining different scores would improve detection. The third is reduced to Eccentricity (Ecc-c), which is weakly correlated with the other displayed scores. It is a black-box graph-based estimator that measures dispersion through semantic contradiction between sampled responses.

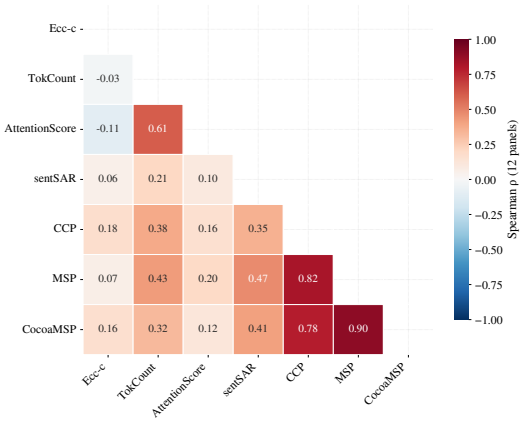


Figure 3: Spearman correlation between the top uncertainty estimators across panels.

Access constraints should shape estimator selection. The ranking clusters differ not only in performance, but also in input requirements. The first one mostly contains white-box estimators that require token probabilities. CocoaMSP and SentenceSAR further require sampled generations and a sentence-similarity model, while AttentionScore requires access to attention weights. In contrast, the NLI-graph estimators are black-box with respect to the generator: they use sampled text responses and an auxiliary NLI model, but do not require generator logits, hidden states, or attention weights. When white-box access is available, CocoaMSP and CCP perform better across datasets, while MSP remains an attractive, computationally efficient baseline.

6 Sensitivity and Controls

In this section, we investigate the sensitivity of our analysis with respect to several factors.

On the correlation between uncertainty estimates and response length. We perform a length-sensitivity analysis on RT, because this is the only dataset with strong heterogeneity of response length, unlike LW and NR. On RT, longer responses hallucinate more often: the hallucination rate rises from 0.27 in the shortest response-length quintile to 0.85 in the longest for L13B. Hence, the number of tokens is a relevant predictor of hallucination. Yet, within quintiles, its performance is close to chance (mean AUROC 0.54 to 0.57). We then test the impact of the response length on the estimators we study. The estimators behave differently with respect to this control. For instance, MSP and CCP remain informative across all five M7B quintiles (0.64 to 0.75) but fall near or below chance on both Llama models. AttentionScore performance improves with length on L13B (0.50 to 0.77 on the different quintiles) while M7B shows no such trend. The complete strata and paired comparisons can be found in Section D.4. We do not find a recurring dependence of uncertainty estimators with response length. Hence, we conclude that the aggregated results presented in Section 5 may hide some length-dependent behavior, yet length by itself does not explain the performance of all estimators.

Impact of the judge model. We perform a sensitivity analysis of our study with respect to the LLM used as a judge to label hallucinations in the HalluLens datasets. Recall that this hallucination label is in $\{0, 1\}$ for NER and PW, while for LW (because of the response length), this labeling is done at the claim-level (a response being divided in several claims), resulting in an averaged score per response in $[0, 1]$. On NER and PW, we find a strong agreement between labeling of the two models, and consequently on the uncertainty estimators results. On LW, where responses are longer, we find a difference of labeling between the two models (with Qwen tagging fewer answers as hallucinated), yet the results of the uncertainty estimators are not affected. This is because the relative ranking of the answers in terms of quality is preserved. Thus, the comparison between uncertainty estimators is stable with respect to the choice of the LLM used as a judge. The protocols and detailed results can be found in Section D.5.

Impact of the hallucination type. In our study, we generally only have the information that a response is hallucinated or not, except on RT, where the dataset additionally incorporates information on the type of hallucination. More specifically, it can be tagged as 'baseless-information' and/or 'evidence-conflict' with respect to the provided evidence (context). We therefore performed a sensitivity analysis on the RT dataset with respect to hallucination type. We find that rankings of the uncertainty estimators' performance remain very similar between the two hallucination types (they are highly correlated), although the best method can differ. Interestingly, the best discrimination performances (the highest AUC) are achieved when considering cases where both subtypes co-occur as the hallucination target. The complete results are reported in Section D.6.

7 Conclusion

Our study clarifies when existing uncertainty estimation methods align with hallucination-related output quality in large language models. The main finding is that uncertainty is informative, but only conditionally. Task-specific evaluation targets are the main source of heterogeneity, which argues against treating any uncertainty estimator as a universal hallucination detector. This leaves the question of which uncertainty estimator should be favored in practice. This choice is actually not only guided by practical use based on the target task and validation performance, but also access to the generator internals and the computational budget.

For the small open-weight models studied here, our results provide practical starting points based on the level of access to the generator. When only sampled outputs are available, contradiction-based graph estimators, including Eccentricity variants, are useful candidates for short-answer correctness and supported long-form quality, but perform less well on context faithfulness and refusal. When token probabilities are also available, MSP provides a strong baseline, while CCP is competitive for short-answer correctness and refusal. When internal attention information is accessible, AttentionScore is a promising candidate for context faithfulness, although its refusal performance is more model-dependent and should be compared against token count. Score correlations further help identify overlapping candidates before investing in more costly methods. These findings guide which estimators to validate, rather than establish a universally best method or a validated ensemble.

Limitations and Future Work

Our conclusions concern three open-weight models and four tasks. In our study, the only dataset with human-annotated hallucinations is RAGTruth. However, this dataset provides these labels only for a fixed set of model responses: MISTRAL7B, LLAMA7B, and LLAMA13B. Hence, because our motivation was to investigate the performance of the widest set of uncertainty estimators across several hallucination settings, we kept this set of models as the evaluated models across alternative hallucination tasks (hallucination datasets). Thus, to conduct a comparative study, we constrained the set of evaluated models. Then, the hallucination tasks we consider cover context faithfulness, short-answer correctness, supported claims in long-form generation, and failure to reject false premises. They do not cover every form of model error.

Future work could extend our evaluation in two directions. First, hallucination targets chould be made more fine-grained. For instance, span-level annotations, such as those available in RAGTruth, can be used to test whether uncertainty scores indicate where hallucinations occur in the answer, rather than only whether a response is unreliable. Second, extension to other datasets and larger and more recent models may strengthen the analysis.

AI use statement

LLMs were used as experimental components to generate question–answer pairs and to evaluate answer quality, claim support, and abstention in the HalluLens tasks, as described in Sections 3.1, C.1 and 4. Separately, generative AI tools helped identify literature during the initial review, understand research papers and code implementations, refactor and review code, improve plot presentation, construct LaTeX tables from CSV/JSON results, and edit prose and mathematical notation. They also helped interpret the reported results and organise the presentation. The authors retain responsibility for the accuracy of the content of the paper.

Ethics statement

This study evaluates hallucination-related failures using existing human-annotated and generated datasets; the reported experiments do not include a new human-annotation study. The principal risk is overinterpreting a benchmark score as evidence that a model response is safe or correct. Uncertainty can miss unsupported content, and LLM judges can introduce systematic errors. Our conclusions are restricted to the evaluated tasks, models, and target definitions, and do not establish reliability in high-stakes applications. We distinguish reference support, answer quality, and abstention so that improvements on one target are not mistaken for a general guarantee against hallucination.

Reproducibility statement

Sections 3.1 and 4 describe the datasets, response models, data splits, logit recovery and evaluation protocol. Section B provides estimator definitions and implementation details; Section C specifies dataset evaluation, reference texts and metric computation. The complementary results include the procedures and evidence for length, judge and target sensitivity (Section D). The code repository is linked in Section B.2. The accompanying code separates heavy GPU-intensive model-dependent score computation from CPU-based evaluation and regeneration of the paper figures and tables, with pinned inputs and configuration files for the latter.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. To Believe or Not to Believe Your LLM: Iterative Prompting for Estimating Epistemic Uncertainty. In *NeurIPS*, November 2024.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863, August 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00881-z. URL https://www.nature.com/articles/s42256-024-00881-z.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Salman Avestimehr, and Sai Praneeth Karimireddy. Reconsidering LLM uncertainty estimation methods in the wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29531–29556, July 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1429. URL https://aclanthology.org/2025.acl-long.1429/.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156, July 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1176. URL https://aclanthology.org/2025.acl-long.1176/.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, Hoda Heidari, Anson Ho, Sayash Kapoor, Leila Khalatbari, Shayne Longpre, Sam Manning, Vasilios Mavroudis, Mantas Mazeika, Julian Michael, Jessica Newman, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Girish Sastry, Elizabeth Seger, Theodora Skeadas, Tobin South, Emma Strubell, Florian Tramèr, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Olubayo Adekanmbi, David Dalrymple, Thomas G. Dietterich, Edward W. Felten, Pascale Fung, Pierre-Olivier Gourinchas, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Andreas Krause, Susan Leavy, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Fahad Albalawi, Marwan Alserkal, Olubunmi Ajala, Guillaume Avrin, Christian Busch, André Carlos Ponce de Leon Ferreira de Carvalho, Bronwyn Fox, Amandeep Singh Gill, Ahmet Halit Hatip, Juha Heikkilä, Gill

Jolly, Ziv Katzir, Hiroaki Kitano, Antonio Krüger, Chris Johnson, Saif M. Khan, Kyoung Mu Lee, Dominic Vincent Ligot, Oleksii Molchanovskyi, Andrea Monti, Nusu Mwamanzi, Mona Nemer, Nuria Oliver, José Ramón López Portillo, Balaraman Ravindran, Raquel Pezoa Rivera, Hammam Riza, Crystal Rugege, Ciarán Seoighe, Jerry Sheehan, Haroon Sheikh, Denise Wong, and Yi Zeng. International AI Safety Report, January 2025. URL http://arxiv.org/abs/2501.17805. arXiv:2501.17805 [cs].

Dylan Bouchard, Mohit Singh Chauhan, David Skarbrevik, Ho-Kyeong Ra, Viren Bajaj, and Zeya Ahmad. UQLM: A Python Package for Uncertainty Quantification in Large Language Models. *Journal of Machine Learning Research*, 27(13):1–10, 2026. ISSN 1533-7928. URL http://jmlr.org/papers/v27/25-1557.html.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=Zj12nzlQbz.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-Level Uncertainty in Large Language Models. In *ICLR*, October 2026. URL https://openreview.net/forum?id=11QZITAMUO.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, NJ, 2001. ISBN 978-0-471-06259-2 978-0-471-20061-1.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pp. 97378–97407, 2024. ISBN 979-8-3313-1438-5.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. Rainproof: An umbrella to shield text generator from out-of-distribution data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5831–5857, 2023.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. ACM, 2006. doi: 10.1145/1143844.1143874.

Siddartha Devic, Tejas Srinivasan, Jesse Thomason, Willie Neiswanger, and Vatsal Sharan. From Calibration to Collaboration: LLM Uncertainty Quantification Should Be More Human-Centered, June 2025. URL http://arxiv.org/abs/2506.07461. arXiv:2506.07461 [cs] version: 1.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5050–5063. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.276. URL https://aclanthology.org/2024.acl-long.276/.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. Lm-polygraph: Uncertainty estimation for language models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In *Findings of the Association for Computational Linguistics*, pp. 9367–9385, August 2024. doi: 10.18653/v1/2024.findings-acl.558. URL https://aclanthology.org/2024.findings-acl.558/.

Dongyang Fan, Sebastien Delsad, Nicolas Flammarion, and Maksym Andriushchenko. HalluHard: A Hard Multi-Turn Hallucination Benchmark, February 2026. URL http://arxiv.org/abs/2602.01031. arXiv:2602.01031 [cs].

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, November 2024a. ISSN 1046-8188. doi: 10.1145/3703155. URL https://dl.acm.org/doi/10.1145/3703155.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in Language Models: Assessment through Rank-Calibration. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 284–312, November 2024b. doi: 10.18653/v1/2024.emnlp-main.18. URL https://aclanthology.org/2024.emnlp-main.18/.

Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. Addressing Pitfalls in the Evaluation of Uncertainty Estimation Methods for Natural Language Generation. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, March 2025. URL https://openreview.net/forum?id=hbIpwrdfE1.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, December 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3571730. URL http://arxiv.org/abs/2202.03629.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, July 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

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.

Sungmin Kang, Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, and Salman Avestimehr. Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions. *arXiv 2510.12040*, 2025. URL https://arxiv.org/abs/2510.12040.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations*, 2023.

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/.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, July 2004. URL https://aclanthology.org/W04-1013/.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Findings of ACL*, 2023.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 10351–10368, November 2024a. doi: 10.18653/v1/2024.emnlp-main.578. URL https://aclanthology.org/2024.emnlp-main.578/.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*, February 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *International Conference on Learning Representations*, 2021.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. Adaptive Retrieval Without Self-Knowledge? Bringing Uncertainty Back Home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6355–6384, July 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.319. URL https://aclanthology.org/2025.acl-long.319/.

Alexander V. 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*, November 2024. URL https://openreview.net/forum?id=j2wCrWmgMX.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, August 2024. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, July 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Xin Qiu and Risto Miikkulainen. Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space. In *Advances in Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=LOH6qzI7T6.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection, June 2021. URL http://arxiv.org/abs/2106.09022. arXiv:2106.09022 [cs].

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11709–11724, November 2024. doi: 10.18653/v1/2024.findings-emnlp.685. URL https://aclanthology.org/2024.findings-emnlp.685/.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. Revisiting Uncertainty Quantification Evaluation in Language Models: Spurious Interactions with Response Length Bias Results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 743–759, July 2025. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.60. URL https://aclanthology.org/2025.acl-short.60/.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In *NeurIPS*, November 2024. URL https://openreview.net/forum?id=LYx4w3CAgy.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, December 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. The Illusion of Certainty: Uncertainty Quantification for LLMs Fails under Ambiguity, January 2026. URL http://arxiv.org/abs/2511.04418. arXiv:2511.04418 [cs].

Liam van der Poel, Ryan Cotterell, and Clara Meister. Mutual Information Alleviates Hallucinations in Abstractive Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5956–5965, December 2022. doi: 10.18653/v1/2022.emnlp-main.399. URL https://aclanthology.org/2022.emnlp-main.399/.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025a. doi: 10.1162/tacl_a_00737. URL https://aclanthology.org/2025.tacl-1.11/.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. CoCoA: A Minimum Bayes Risk Framework Bridging Confidence and Consistency for Uncertainty Quantification in LLMs. In *NeurIPS*, October 2025b. URL https://openreview.net/forum?id=H1NGlLNaVC.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11659–11681, July 2023. doi: 10.18653/v1/2023.acl-long.652. URL https://aclanthology.org/2023.acl-long.652/.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. Uncertainty-Aware Attention Heads: Efficient Unsupervised Uncertainty Quantification for LLMs. *arXiv 2505.20045*, 2025. doi: 10.48550/ARXIV.2505.20045. URL https://arxiv.org/abs/2505.20045.

Xunzhi Wang, Zhuowei Zhang, Gaonan Chen, Qiongyu Li, Bitong Luo, Zhixin Han, Haotian Wang, Zhiyu Li, Hang Gao, and Mengting Hu. UBench: Benchmarking Uncertainty in Large Language Models with Multiple Choice Questions. In *Findings of the Association for Computational Linguistics*, pp. 8076–8107, July 2025. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.423. URL https://aclanthology.org/2025.findings-acl.423/.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, November 2024. URL http://arxiv.org/abs/2411.04368. arXiv:2411.04368 [cs].

Ronald J. Williams and David Zipser. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280, June 1989. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1989.1.2.270. URL https://direct.mit.edu/neco/article/1/2/270-280/5490.

Xianzong Wu, Xiaohong Li, Lili Quan, and Qiang Hu. UncertaintyZoo: A Unified Toolkit for Quantifying Predictive Uncertainty in Deep Learning Systems, December 2025. URL http://arxiv.org/abs/2512.06406. arXiv:2512.06406 [cs].

Qi Yang, Shreya Ravikumar, Fynn Schmitt-Ulms, Satvik Lolla, Ege Demir, Iaroslav Elistratov, Alex Lavaee, Sadhana Lolla, Elaheh Ahmadi, Daniela Rus, Alexander Amini, and Alejandro Perez. Uncertainty-aware Language Modeling for Selective Question Answering, November 2023. URL http://arxiv.org/abs/2311.15451. arXiv:2311.15451 [cs].

Yuxuan Yao, Han Wu, Zhijiang Guo, Zhou Biyan, Jiahui Gao, Sichun Luo, Hanxu Hou, Xiaojin Fu, and Linqi Song. Learning From Correctness Without Prompting Makes LLM Efficient Reasoner. In *COLM*, August 2024. URL https://openreview.net/forum?id=dcbNzhVVQj#discussion.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. Detection of Adversarial Examples in Text Classification: Benchmark and Baseline via Robust Density Estimation. In *Findings of the Association for Computational Linguistics*, pp. 3656–3672, May 2022. doi: 10.18653/v1/2022.findings-acl.289. URL https://aclanthology.org/2022.findings-acl.289/.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. LUQ: Long-text Uncertainty Quantification for LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 5244–5262, November 2024. doi: 10.18653/v1/2024.emnlp-main.299. URL https://aclanthology.org/2024.emnlp-main.299/.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, March 2025. URL http://arxiv.org/abs/2303.18223. arXiv:2303.18223 [cs].

APPENDIX CONTENTS

A. Related works 16
B. Uncertainty estimators 17
C. Additional methodology 30
Datasets and response evaluation 30
AlignScore reference texts 32
Evaluation metrics 32
D. Additional results 32
Complete performance tables 32
Agreement between estimator rankings 37
Performance by task and estimator family 40
Response length 44
Changing the judge 46
RAGTruth: error types and proportions 47
LongWiki: precision, recall and error thresholds 50

A Related works

A.1 Evaluating uncertainty estimation methods

Work on evaluating uncertainty estimation (UE) methods for LLMs has addressed three concerns: building infrastructure that places heterogeneous methods on a common framework; assessing whether comparisons drawn under controlled conditions hold up in deployment; and questioning whether the evaluation protocols themselves are sound. Within the third concern, we examine whether the evaluation target reflects the particular response failure that uncertainty is intended to detect.

Method-coverage frameworks. The most direct response to method heterogeneity has been to consolidate estimators behind shared interfaces and shared intermediate quantities. LM-Polygraph (Vashurin et al., 2025a) implements a comprehensive estimator suite behind a single pipeline that computes greedy responses, token probabilities, sampled responses, and semantic relation matrices once and reuses them across estimators, controlling for implementation differences. Adjacent toolkits – UQLM (Bouchard et al., 2026), UncertaintyZoo (Wu et al., 2025), and UBench (Wang et al., 2025) – provide complementary coverage or standardize estimator outputs to a common confidence range. Across these frameworks, the contribution is comparability rather than validity: each evaluation is only as meaningful as its target for correctness.

Deployment- and protocol-level critiques. A second strand asks whether rankings produced under controlled benchmarks transfer to deployment, and whether the protocols themselves can be trusted. Bakman et al. (2025) evaluate nineteen UE methods under decision-threshold drift, prompt perturbation, long-form adaptation, and ensembling, and find that most methods are sensitive to threshold choice and vulnerable to adversarial prompts – patterns invisible under standard evaluation conditions. Santilli et al. (2025) show that the same response length can bias both UE scores and correctness functions so that a trivial length baseline can rival principled methods under length-biased correctness metrics. Ielanskyi et al. (2025) reinforce this concern by showing that the choice of approximate correctness function (ROUGE, BLEURT, LLM-as-judge variants) substantially reorders method rankings, and propose marginalising over multiple judges as a more robust alternative. Complementary critiques target the limited ecological validity of QA-style evaluations (Devic et al., 2025), the inability of realisable benchmarks to test calibration over inherently variable outcomes (Cruz et al., 2024; Tomov et al., 2026), and the incomparable scales on which estimator outputs live (Huang et al., 2024b). The shared message is that "method A outperforms method B on benchmark X" is routinely confounded by length, choice of correctness function, threshold calibration, or estimator output scale – and that fixing the benchmark itself does not fix the comparison.

The evaluation gap we address. Across these strands, a common limitation is that the evaluation *target* is rarely a hallucination-specific signal: correctness is operationalised through lexical or semantic overlap with a reference answer, or through a multiple-choice answer key, rather than through an explicit characterisation of when a response is unsupported by the evidence the model was supposed to use. Our contribution sits in this space: we retain the LM-Polygraph estimator catalogue and the methodological caution of Santilli et al. (2025), Bakman et al. (2025), and Ielanskyi et al. (2025), but replace generic correctness targets with hallucination targets defined through operational distinctions established in the hallucination literature, which we develop next.

A.2 Hallucination: definitions and benchmarks

The definition of hallucination has shifted as LLMs have moved from specific tasks to largely open-ended generation, and the choice of definition directly shapes what a benchmark measures.

From task-specific taxonomies to LLM-specific definitions. The intrinsic/extrinsic distinction was originally formulated for specific natural language generation tasks by Ji et al. (2023): a response is *intrinsically* hallucinated if it contradicts an input source, and *extrinsically* hallucinated if it cannot be verified from that source – a notion that presupposes a single explicit input. As LLMs increasingly generate free-form text without such a source, two reframings followed. Huang et al. (2024a) proposed *factuality* versus *faithfulness* as the LLM-era replacement axis (correctness against the world vs. adherence to context), while Augenstein et al. (2024) argued that *factuality* and *hallucination* should

be treated as distinct concepts entirely: factuality requires an external oracle, whereas hallucination admits an internal definition – whether the response matches the evidence the model was supposed to use. The separation is consequential because the two concepts overlap but are not identical: a model can be factually correct while unfaithful to its evidence, or factually wrong while faithful to a misleading source.

HalluLens and the new extrinsic/intrinsic taxonomy. Bang et al. (2025) consolidate this picture and introduce the HALLULENS benchmark. They redefine *intrinsic* hallucinations as responses unsupported by information *explicitly provided at inference time*, and *extrinsic* hallucinations as responses unsupported by the *pretraining knowledge the model is expected to have encountered*. The reformulation is operational: each regime corresponds to a concrete source of evidence against which a response can be evaluated, with ground-truth labels defined without implicit appeal to world truth. HALLULENS instantiates this with three extrinsic tasks anchored to Wikipedia, along with items generated on-the-fly to mitigate test-set leakage. The Wikipedia anchor rests on a documented approximation – recent pretraining mixtures are assumed to contain Wikipedia, so inconsistency with a Wikipedia article is treated as a proxy for inconsistency with pretraining knowledge and remains imperfect for long-form generation (Fan et al., 2026).

Benchmark heterogeneity and our positioning. With the operational split in hand, the existing landscape can be repartitioned by the definition each benchmark encodes. *Explicitly intrinsic* benchmarks include RAGTRUTH (Niu et al., 2024), which provides word-level hallucination annotations for retrieval-augmented generation. Closed-book QA benchmarks such as TRIVIAQA (Joshi et al., 2017), NATURAL QUESTIONS (Kwiatkowski et al., 2019), and SIMPLEQA (Wei et al., 2024) target some dimension of factuality but sit ambiguously with respect to the operational split: classifying them as extrinsic-regime benchmarks would require evidence that all their questions are answerable from the model's pretraining data, which is generally not established. HALLULENS fills this gap by restricting questions to those generated from Wikipedia, making the pretraining-coverage assumption a documented approximation rather than an implicit one; *explicitly extrinsic* benchmarks otherwise remain comparatively rare. As a result, comparing performance across these benchmarks partly compares definitions rather than methods. We therefore pair RAGTRUTH (intrinsic regime, with human span annotations) with the three HALLULENS tasks – PRECISEWIKIQA, LONGWIKI, and NONEXISTENTREFUSAL – to cover three extrinsic settings with explicit evaluation procedures. Together, these commit us to the hallucination/factuality distinction articulated by Augenstein et al. (2024) and operationalized by Bang et al. (2025), and allow us to compare uncertainty methods while keeping the meaning of each target explicit. Cross-regime differences still combine target, task, and generation effects.

B DESCRIPTION AND IMPLEMENTATION DETAILS ON UNCERTAINTY ESTIMATORS

This appendix specifies the uncertainty estimators evaluated in our benchmark, briefly described in Section 2. We pursue two goals jointly: a conceptual map organized by the signal each method exploits, and an implementation-faithful description of the exact variants executed in our pipeline. Every formula below was derived by auditing the corresponding implementation in our LM-POLYGRAPH-based pipeline (Vashurin et al., 2025a; Fadeeva et al., 2023) rather than from the high-level descriptions in the original papers.

Pipeline. We use LM-POLYGRAPH as the uncertainty-score computation engine. The framework separates two stages: shared intermediate quantities (greedy responses, token log-probabilities, stochastic samples, semantic relation matrices, internal representations) are computed *once* per input, and the estimators consume these quantities to produce the final scores. We do not modify any estimator implementation. The version of LM-POLYGRAPH we use is `main/0e5bcdd35efe7de4e3ae8124427aa12575a1bd2c`. Downstream evaluation is performed in our analysis pipeline, externally to LM-POLYGRAPH.

B.1 Implementation details

Table 2 consolidates the numerical hyperparameters and external models used in the executed pipeline. Default LM-Polygraph settings are used for any value not listed.

B.2 Code availability

The code is available here: https://anonymous.4open.science/r/uncertainty-benchmark-867C.

B.3 Compute resources

All GPU-demanding computation was performed on a university HPC cluster using usage-constrained GPU allocations. The primary hardware was an NVIDIA RTX PRO 6000 Blackwell (96 GB GDDR7) for large-model runs, with NVIDIA RTX 6000 Ada/RTX A6000 (48 GB GDDR6) nodes used for smaller models. Nodes were equipped with 24–32 CPU cores and 192–512 GB RAM. The two main GPU-demanding phases totaled approximately 700 productive GPU-hours:

- Benchmark inference (3 tasks × 3 models, up to 2,000 samples each, 50 uncertainty estimators): roughly 415 GPU-hours for HalluLens tasks and 255 GPU-hours for RAGTruth, ranging from 5 GPU-hours per run (small datasets, 7B models) to 87 GPU-hours (large datasets, 13B models). 48 GB of VRAM suffices for 7B models at a batch size of 1; 13B models require 96 GB.
- LongWiki evaluation (Llama-3.1-70B-AWQ judge via vLLM): 31 productive GPU-hours across 2 model runs, using 2 GPUs simultaneously (judge server + evaluation client).

B.4 Notation and conventions

Let x denote the input query and $\hat{y} = (\hat{y}_1, \dots, \hat{y}_T)$ the *response to be scored* of length T. In the standard free-generation setting, $\hat{y}$ is the greedy response of the model; in the RAGTruth dataset, where responses have already been generated and annotated, $\hat{y}$ is the observed response supplied by the dataset. The autoregressive conditional probability of the token $\hat{y}_t$ is $p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t})$, and the unconditional language-model probability (input x removed) is $p_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{y}_{<t})$. The full predictive distribution at position t lives on the simplex $\Delta(\mathcal{V})$ over the vocabulary of size $V = |\mathcal{V}|$, i.e. $p_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<t}) \in \Delta(\mathcal{V})$. The full sequence probability is $P_{\boldsymbol{\theta}}(\hat{y} \mid x) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t})$. Shannon entropy of any $q \in \Delta(\mathcal{V})$ is $\mathcal{H}(q) = -\sum_{v \in \mathcal{V}} q(v) \log q(v)$, and $\mathcal{U}(\mathcal{V})$ denotes the uniform distribution on $\mathcal{V}$.

When stochastic sampling is required, we draw $S = 10$ responses $y^{(1)}, \dots, y^{(S)} \sim p_{\boldsymbol{\theta}}(\cdot \mid x)$ at sampling temperature 1 with random seed 42. We write $|y^{(s)}|$ for the length of sample s and $P_{\boldsymbol{\theta}}(y^{(s)} \mid x) = \prod_{t=1}^{|y^{(s)}|} p_{\boldsymbol{\theta}}(y_t^{(s)} \mid x, y_{<t}^{(s)})$ for its sequence probability. The same sample pool is shared across all sample-based estimators for a given input, so any two such estimators are compared on identical samples. For semantic-consistency methods, NLI-based pairwise scores are

$$E_{ij} = p_{\text{NLI}}(\text{entailment} \mid y^{(i)}, y^{(j)}), \quad C_{ij} = p_{\text{NLI}}(\text{contradiction} \mid y^{(i)}, y^{(j)}),$$

produced by `microsoft/deberta-large-mnli` (He et al., 2021).[3] Cross-encoder semantic similarity, when used, is computed with `cross-encoder/stsb-roberta-large`.[4]

Score orientation. All estimators are reported as *uncertainty* scores: larger values indicate greater uncertainty. Two estimators in LM-Polygraph are confidence-oriented (Rényi Divergence and Fisher–Rao Distance) and are negated before analysis. The formulas below are uncertainty-oriented.

Name versus executed formula. Two estimator names refer to a probability, whereas the executed scores are log-transformations of that probability. We make this explicit to avoid confusion. Maximum Sequence Probability refers to the quantity $P_{\boldsymbol{\theta}}(\hat{y} \mid x)$, but the implemented score in LM-Polygraph is its negative log – the sequence negative log-likelihood (NLL). Similarly, Perplexity refers to

[3] https://huggingface.co/microsoft/deberta-large-mnli

[4] https://huggingface.co/cross-encoder/stsb-roberta-large

$\exp\big(-\frac{1}{T}\sum_t \log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t})\big)$, whereas the executed score is the *logarithm* of this perplexity, i.e. the length-normalized sequence negative log-likelihood (LN-NLL).

Definitional convention. We use $:=$ for the exact quantity returned by the implementation and $=$ for subsequent algebraic rewrites.

Configuration convention. Unless stated otherwise, we use the default LM-POLYGRAPH configuration. When a hyperparameter materially affects the score, we report its value, see Table 2.

B.5 WHITE-BOX ESTIMATORS

White-box estimators require access to information beyond the generated text of the response: the model's logits, attention maps, hidden states, or training data. We organize them into five families, ordered from the most direct readouts of the predictive distribution to the most indirect: (i) information-based estimators, which aggregate token-level distributional quantities computed in a single forward pass; (ii) sample-dispersion estimators, which lift these quantities to multiple stochastic samples and combine them with measures of inter-sample dispersion; (iii) internal-state estimators, which use attention or hidden-state geometry; (iv) training-based estimators, which score representation-space distance from a fitted reference distribution; and (v) reflexive estimators, which prompt the model to judge its own answer.

B.5.1 INFORMATION/LOGIT-BASED ESTIMATORS

These estimators consume only the token-level predictive distributions encountered along a single response $\hat{y}$ and aggregate a position-wise quantity across the sequence. Within this family we distinguish methods that look only at the probability of the realized token (Maximum Sequence Probability, Perplexity), methods that examine the full predictive distribution at each position (Mean Token Entropy, SelfCertainty, Rényi Divergence, Fisher–Rao Distance), methods that contrast the conditional predictive distribution with its unconditional counterpart (Pointwise Mutual Information and its conditional variant), and two methods that introduce auxiliary signals at the token level (TokenSAR, Claim-Conditioned Probability).

Maximum Sequence Probability. Maximum Sequence Probability is the most basic logit-based estimator: it considers only the probability assigned to the realized response and ignores the rest of the predictive distribution. The executed score is the sequence negative log-likelihood,

$$U_{\mathrm{MSP}}(x;\hat{y}) \;:=\; -\sum_{t=1}^{T} \log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t}) \;=\; -\log P_{\boldsymbol{\theta}}(\hat{y} \mid x).$$

The score is zero when the model assigns probability 1 to every realized token and grows without bound as confidence in the realized response drops. It is length-sensitive: longer responses accumulate more uncertainty at constant per-token confidence.

Perplexity. Perplexity removes the length sensitivity of Maximum Sequence Probability by averaging the per-token negative log-likelihood,

$$U_{\mathrm{PPL}}(x;\hat{y}) \;:=\; -\frac{1}{T}\sum_{t=1}^{T} \log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t}),$$

which equals the logarithm of the standard perplexity of $\hat{y}$ under the model. Like Maximum Sequence Probability, this score uses only the probability of the realized token, but it is comparable across responses of different lengths (Fomicheva et al., 2020).

Mean Token Entropy. Mean Token Entropy is the simplest logit-based estimator that uses the *full* predictive distribution at each position rather than only the probability of the realized token. The score averages the Shannon entropy of the predictive distribution across positions, which equals the chain-rule expansion of the sequence entropy (Cover & Thomas, 2001; Malinin & Gales, 2021):

$$U_{\mathrm{MTE}}(x) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \mathcal{H}\big(p_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<t})\big).$$

Unlike MSP and PPL, MTE is not determined by the realized token alone: even when $\hat{y}_t$ is itself among the most probable tokens, MTE can still be large if probability mass is spread broadly across alternatives; conversely, when $\hat{y}_t$ has only moderate probability but the distribution is peaked on a few competing tokens, MTE remains small because the distribution is concentrated.

SelfCertainty. SelfCertainty compares a uniform reference distribution to the model's predictive distribution, using the *reverse* KL direction relative to the entropy-like divergence-from-uniform scores used by Rényi Divergence and Fisher–Rao Distance below. Let $\tilde{\mathcal{V}}_t \subseteq \mathcal{V}$ be the set of vocabulary entries with finite log-probability at position t. The implemented token-level quantity is

$$D_{\mathrm{KL}}\big(\mathcal{U}(\tilde{\mathcal{V}}_t) \,\|\, p_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<t})\big) = -\log|\tilde{\mathcal{V}}_t| - \frac{1}{|\tilde{\mathcal{V}}_t|}\sum_{v\in\tilde{\mathcal{V}}_t} \log p_{\boldsymbol{\theta}}(v \mid x, \hat{y}_{<t}),$$

and the sequence-level uncertainty score reverses the polarity by taking the negative mean:

$$U_{\mathrm{SC}}(x) \;:=\; -\frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\big(\mathcal{U}(\tilde{\mathcal{V}}_t) \,\|\, p_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<t})\big).$$

A flat predictive distribution yields a small divergence and $U_{\mathrm{SC}} \approx 0$ (high uncertainty); a sharply peaked distribution yields a large divergence and $U_{\mathrm{SC}} \ll 0$ (low uncertainty). Unlike entropy or Rényi-style scores, the reverse-KL direction is especially sensitive to tokens assigned very small probability: a single very small $p_{\boldsymbol{\theta}}(v \mid \cdot)$ blows up one term of the sum.

Rényi Divergence. Rényi Divergence also compares the predictive distribution to a uniform reference, but in the *opposite direction* from SelfCertainty – and on a temperature-scaled version of the distribution. Let

$$q_t(v \mid x, \hat{y}_{<t}; \tau) = \mathrm{softmax}(\log p_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<t}) \,/\, \tau)(v) \;\in\; \Delta(\mathcal{V}).$$

The token-level Rényi divergence of order α from q_t to the uniform distribution is

$$D_\alpha\big(q_t(\cdot \mid x, \hat{y}_{<t}; \tau) \,\|\, \mathcal{U}(\mathcal{V})\big) = \frac{1}{\alpha - 1}\log\sum_{v\in\mathcal{V}} q_t(v \mid x, \hat{y}_{<t}; \tau)^\alpha - \log|\mathcal{V}|,$$

and the sequence-level uncertainty score, after polarity reversal, is

$$U_{\mathrm{Renyi}}(x; \tau) \;:=\; -\frac{1}{T}\sum_{t=1}^{T} D_\alpha\big(q_t(\cdot \mid x, \hat{y}_{<t}; \tau) \,\|\, \mathcal{U}(\mathcal{V})\big).$$

Larger divergence corresponds to a sharper, more concentrated distribution, hence the negation. The parameter α controls sensitivity to the distribution: $\alpha < 1$ gives more weight to low-probability tokens, $\alpha > 1$ emphasizes the peak; the temperature τ smooths the distribution before the divergence is computed. SelfCertainty should *not* be read as the $\alpha = 1$ case of this estimator: it uses the KL direction $\mathcal{U} \,\|\, p$, whereas this Rényi score uses $q \,\|\, \mathcal{U}$, which measures a different geometric quantity (peak concentration rather than coverage). We use $\alpha = 0.5$ and $\tau = 2$ (Darrin et al., 2023).

Fisher–Rao Distance. Fisher–Rao Distance also measures how far the temperature-scaled predictive distribution $q_t(\cdot \mid x, \hat{y}_{<t}; \tau)$ lies from the uniform distribution on $\Delta(\mathcal{V})$, but uses Fisher–Rao geometry rather than a divergence. The token-level geodesic distance is

$$d_{\mathrm{FR}}\big(q_t(\cdot \mid x, \hat{y}_{<t}; \tau), \mathcal{U}(\mathcal{V})\big) = \frac{2}{\pi}\arccos\left(\sum_{v\in\mathcal{V}}\sqrt{q_t(v \mid x, \hat{y}_{<t}; \tau)\cdot\tfrac{1}{|\mathcal{V}|}}\right),$$

and the sequence-level uncertainty score, after polarity reversal, is

$$U_{\mathrm{FR}}(x; \tau) \;:=\; -\frac{1}{T}\sum_{t=1}^{T} d_{\mathrm{FR}}\big(q_t(\cdot \mid x, \hat{y}_{<t}; \tau), \mathcal{U}(\mathcal{V})\big).$$

The geodesic distance is zero when q_t is uniform and approaches 1 as q_t becomes more peaked; the negation orients the score so that a flatter (more uncertain) distribution receives a larger value. Unlike any divergence, Fisher-Rao is a true metric on the simplex (symmetric, satisfies the triangle inequality). We use $\tau = 2$ (Darrin et al., 2023).

Mean Pointwise Mutual Information. The estimators above measure either how confident the model is in the realized response (MSP, PPL) or how broadly its predictive distribution spreads (MTE, SelfCertainty, Rényi, Fisher–Rao). They cannot detect responses that are likely under the model regardless of the input, that is, responses whose probability does not benefit from prompt grounding. Mean Pointwise Mutual Information addresses this by contrasting, for each realized token, the conditional log-probability $\log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t})$ with the unconditional log-probability $\log p_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{y}_{<t})$. The latter is obtained by an additional forward pass on the response alone, with the prompt x removed. The executed score is

$$U_{\mathrm{PMI}}(x; \hat{y}) \;:=\; -\frac{1}{T}\sum_{t=1}^{T}\big[\log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t}) - \log p_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{y}_{<t})\big].$$

When the conditional probability exceeds the unconditional, the prompt is informative and U_{PMI} is negative (low uncertainty); when the two are comparable, the response could have been generated without the prompt and U_{PMI} approaches zero (high uncertainty), which is interpreted as a signal of weak grounding (van der Poel et al., 2022).

Mean Conditional Pointwise Mutual Information. Mean Conditional Pointwise Mutual Information refines PMI by gating the unconditional correction: the correction is applied only at positions where the predictive distribution is locally uncertain, and is ignored at sharp positions, where the unconditional log-probability would add noise rather than signal. "Locally uncertain" is operationalized through the same per-position entropy used by MTE: the correction is applied only when $\mathcal{H}(p_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<t})) \geq \tau$, with weight λ. The executed score is

$$U_{\mathrm{CPMI}}(x; \hat{y}) \;:=\; -\frac{1}{T}\sum_{t=1}^{T}\Big[\log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t}) - \lambda \cdot \mathbf{1}\big[\mathcal{H}(p_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<t})) \geq \tau\big] \cdot \log p_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{y}_{<t})\Big].$$

At low-entropy positions, the score reduces to the realized-token negative log-likelihood (Perplexity-like behavior); at high-entropy positions, it behaves like PMI (van der Poel et al., 2022). We use the LM-POLYGRAPH defaults $\tau = 0.0656$ and $\lambda = 3.599$.

TokenSAR. TokenSAR re-weights the per-token negative log-likelihoods by an external measure of token relevance, so that tokens whose removal changes the meaning of the response contribute more to the score than function words or filler. Token relevance is computed by leave-one-out cross-encoder similarity: for each position t, let $\mathrm{sim}_t \in [0, 1]$ be the semantic similarity between $\hat{y}$ and the response with token $\hat{y}_t$ removed. The unnormalized relevance is $1 - \mathrm{sim}_t$, and the normalized relevance is $R_t = (1 - \mathrm{sim}_t)/\sum_{t'=1}^{T}(1 - \mathrm{sim}_{t'})$. The executed score is

$$U_{\mathrm{TSAR}}(x; \hat{y}) \;:=\; \sum_{t=1}^{T} R_t \cdot \big[-\log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t})\big].$$

Only high-relevance improbable tokens drive uncertainty; an improbable but semantically dispensable token has little effect (Duan et al., 2024).

Claim-Conditioned Probability. Claim-Conditioned Probability targets a different failure mode of standard token-probability estimators: at each position, several alternative tokens may be plausible substitutes, and only some of them would change the meaning of the response. The executed pipeline proceeds in three steps. At each position t, the top-10 token alternatives $\mathcal{A}_t = \{a_{t,0}, a_{t,1}, \ldots, a_{t,9}\}$ are identified, where $a_{t,0}$ is the observed token if it belongs to the selected alternatives. The post-hoc calculator moves that token to the first position when present; it does not insert it when it lies outside the top ten. Each alternative is decoded to text and stripped of surrounding punctuation and whitespace. Bidirectional NLI compares these token strings with the first alternative, rather than comparing complete substituted responses. The two directional labels are combined into a single label $c_{t,i} \in \{\mathrm{entail}, \mathrm{contra}, \mathrm{neutral}\}$: if both directions agree, that label is used; if one direction is entail and the other contra, the label is neutral; otherwise the single non-neutral label is used. Alternatives with $c_{t,i} = \mathrm{entail}$ and the first alternative $a_{t,0}$ (included by construction) form the entailment set $\mathcal{E}_t$, alternatives with $c_{t,i} = \mathrm{contra}$ form the contradiction set $\mathcal{C}_t$, and neutral alternatives are discarded. At

position t, the claim-conditioned probability is the share of probability mass on entailing alternatives within the entailing-or-contradicting subset,

$$\mathrm{CCP}_t = \frac{\sum_{a \in \mathcal{E}_t} p_{\boldsymbol{\theta}}(a \mid x, \hat{y}_{<t})}{\sum_{a \in \mathcal{E}_t \cup \mathcal{C}_t} p_{\boldsymbol{\theta}}(a \mid x, \hat{y}_{<t})},$$

and the sequence-level uncertainty score is the negated product across positions,

$$U_{\mathrm{CCP}}(x) \;:=\; -\prod_{t=1}^{T} \mathrm{CCP}_t.$$

Since neutral alternatives are excluded from the denominator, the ratio measures semantic support *conditional on the alternative being either entailing or contradicting*, not support against the full predictive distribution. For a supplied RAGTruth response, the first alternative need not be the observed token; the frequency and effect of this case have not been quantified. A single position at which probability leaks toward contradicting alternatives is enough to pull the product toward zero and the score toward zero in absolute value (high uncertainty); concentration of mass on entailing alternatives produces a near-1 product and a score near -1 (low uncertainty) (Fadeeva et al., 2024).

B.5.2 Sample semantic dispersion estimators

Sample-dispersion estimators draw S stochastic samples $y^{(1)}, \ldots, y^{(S)}$ from the predictive distribution and combine sample-level probabilities with a measure of dispersion across samples. Beyond the probabilities themselves, several estimators in this family use auxiliary signals (semantic relations from an NLI model, semantic similarity from a cross-encoder, or internal embeddings); we characterize them as probability-aware sample-based estimators rather than as pure consistency estimators. They mirror the structure of the logit-based family: probability-only methods (Monte Carlo Sequence Entropy and its length-normalized variant), methods that summarize agreement under a notion of meaning (Semantic Entropy, Semantic Density), and hybrid methods that combine probability with an external relevance or consistency signal (SentenceSAR, SAR, the Cocoa family).

Monte Carlo Sequence Entropy. Monte Carlo Sequence Entropy approximates the model's expected sequence negative log-likelihood under its own sampling distribution:

$$U_{\mathrm{MCSE}}(x) \;:=\; -\frac{1}{S}\sum_{s=1}^{S} \log P_{\boldsymbol{\theta}}(y^{(s)} \mid x) \;=\; -\frac{1}{S}\sum_{s=1}^{S}\sum_{t=1}^{|y^{(s)}|} \log p_{\boldsymbol{\theta}}\big(y_t^{(s)} \mid x, y_{<t}^{(s)}\big).$$

A model that consistently produces high-probability samples receives a low score; a model that scatters probability mass across many possible responses receives a high score (Malinin & Gales, 2021).

Monte Carlo Normalized Sequence Entropy. Monte Carlo Normalized Sequence Entropy is the length-normalized version of the previous estimator: each sample's negative log-likelihood is divided by its own length before averaging, so that the score is comparable across prompts that elicit responses of different lengths:

$$U_{\mathrm{MCNSE}}(x) \;:=\; -\frac{1}{S}\sum_{s=1}^{S} \frac{1}{|y^{(s)}|} \log P_{\boldsymbol{\theta}}(y^{(s)} \mid x).$$

Semantic Entropy. Semantic Entropy uses the same sample-based negative log-likelihoods, but groups them by meaning before computing entropy, so that paraphrases of the same answer do not inflate the score. The pipeline first partitions the samples into semantic equivalence classes $C_1, \ldots, C_M$ using bidirectional NLI entailment: $y^{(i)}$ and $y^{(j)}$ are placed in the same class when they entail one another. The probability of class C_m is then obtained by summing sample probabilities and renormalizing:

$$P(C_m \mid x) = \frac{\sum_{s\,:\,y^{(s)} \in C_m} P_{\boldsymbol{\theta}}(y^{(s)} \mid x)}{\sum_{s=1}^{S} P_{\boldsymbol{\theta}}(y^{(s)} \mid x)}.$$

The uncertainty score is the entropy of this class distribution:

$$U_{\mathrm{SE}}(x) \;:=\; -\sum_{m=1}^{M} P(C_m \mid x) \log P(C_m \mid x).$$

A model that always produces the same meaning regardless of surface form yields a single class and zero entropy; a model that distributes mass over several meanings yields large entropy (Kuhn et al., 2023; Farquhar et al., 2024).

Semantic Density. Semantic Density replaces the hard equivalence classes of Semantic Entropy with a soft NLI-based kernel and a probability-weighted aggregation around the greedy response $\hat{y}$. Let $\tilde{P}_{\boldsymbol{\theta}}(y^{(s)} \mid x) = \exp\!\big(\frac{1}{|y^{(s)}|} \log P_{\boldsymbol{\theta}}(y^{(s)} \mid x)\big)$ denote the length-normalized sample probability, and let K_s denote the NLI-based kernel score that quantifies the soft semantic agreement between sample s and the surrounding samples (including $\hat{y}$). Writing $\tilde{P}_{\boldsymbol{\theta}}(\hat{y} \mid x)$ for the length-normalized probability of the greedy response, the executed score is the negated probability-weighted kernel density

$$U_{\mathrm{SD}}(x; \hat{y}) \;:=\; -\frac{\sum_{s=1}^{S} \tilde{P}_{\boldsymbol{\theta}}(y^{(s)} \mid x)\, K_s + \tilde{P}_{\boldsymbol{\theta}}(\hat{y} \mid x)}{\sum_{s=1}^{S} \tilde{P}_{\boldsymbol{\theta}}(y^{(s)} \mid x) + \tilde{P}_{\boldsymbol{\theta}}(\hat{y} \mid x)}.$$

The score is close to -1 (low uncertainty) when the probable samples cluster in a dense semantic neighborhood of $\hat{y}$, and close to 0 (high uncertainty) when the kernel scores are uniformly low (Qiu & Miikkulainen, 2024).

SentenceSAR. SentenceSAR extends the relevance-weighting idea of TokenSAR to the multi-sample setting. The contribution of each sample to the score is its own probability augmented by a relevance term that aggregates the probabilities of *other* samples weighted by their semantic similarity to the current one. Writing sim_{sj} for the cross-encoder similarity between $y^{(s)}$ and $y^{(j)}$ and τ for a temperature parameter, the support term is

$$R_S(s; \tau) = \frac{1}{\tau} \sum_{j \neq s} P_{\boldsymbol{\theta}}(y^{(j)} \mid x)\, \mathrm{sim}_{sj},$$

and the executed score is

$$U_{\mathrm{SentSAR}}(x; \tau) \;:=\; \frac{1}{S} \sum_{s=1}^{S} -\log\!\big(P_{\boldsymbol{\theta}}(y^{(s)} \mid x) + R_S(s; \tau)\big).$$

A sample that is both probable and semantically supported by other samples contributes little to the score; a sample that is improbable and semantically isolated contributes a large negative log (Duan et al., 2024). We use $\tau = 1$.

SAR. SAR is the full pipeline of Duan et al. (2024), combining TokenSAR at the token level and SentenceSAR at the sentence level. The raw sample probability used in SentenceSAR is replaced with the relevance-adjusted probability from TokenSAR. Concretely, with $\tilde{P}^{\mathrm{tsar}}_{\boldsymbol{\theta}}(y^{(s)}) = \exp\!\big(-U_{\mathrm{TSAR}}(x, y^{(s)})\big)$ the relevance-weighted probability of sample s, the support term becomes $\tilde{R}_S(s; \tau) = (1/\tau) \sum_{j \neq s} \tilde{P}^{\mathrm{tsar}}_{\boldsymbol{\theta}}(y^{(j)})\, \mathrm{sim}_{sj}$, and the executed score is

$$U_{\mathrm{SAR}}(x; \tau) \;:=\; \frac{1}{S} \sum_{s=1}^{S} -\log\!\big(\tilde{P}^{\mathrm{tsar}}_{\boldsymbol{\theta}}(y^{(s)}) + \tilde{R}_S(s; \tau)\big).$$

SAR composes two relevance signals – token-level for each sample and sentence-level across samples – into a single uncertainty score. We use $\tau = 1$.

Cocoa family. The Cocoa family follows a simple template: a base uncertainty score is multiplied by an inter-sample semantic dissimilarity factor, so that the final score is large only when the response is uncertain on its own *and* semantically dissimilar to the alternative samples (Vashurin et al., 2025b). Let $u_{\mathrm{base}}(x; \hat{y})$ denote the base uncertainty and let $\boldsymbol{S}^{\mathrm{sent}} \in \mathbb{R}^{K \times K}$ be the cross-encoder sentence-similarity matrix over the K sentence units computed by LM-POLYGRAPH for this input (greedy

response and/or samples, depending on the module configuration). The dissimilarity factor is the mean of $1 - S_{ij}^{\text{sent}}$ across the full matrix,

$$D_{\text{sent}}(x) = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \big(1 - S_{ij}^{\text{sent}}\big),$$

and the executed Cocoa score is the product

$$U_{\text{Cocoa}}(x; \hat{y}) \;:=\; u_{\text{base}}(x; \hat{y}) \cdot D_{\text{sent}}(x).$$

We instantiate three variants by choosing $u_{\text{base}} \in \{U_{\text{MSP}}, U_{\text{PPL}}, U_{\text{MTE}}\}$, denoted CocoaMSP, CocoaPPL, and CocoaMTE. An improbable response that is semantically consistent with the alternative samples is treated as a paraphrase and receives a low score; an improbable response that is also semantically isolated receives a high score. *Implementation note:* although the LM-POLYGRAPH source comment states that the diagonal is excluded, the executed code averages over the full $K \times K$ matrix. Since self-similarity is 1, the diagonal entries contribute 0 to $1 - S_{ij}^{\text{sent}}$ and thus dilute D_{sent} by a factor of $(K-1)/K$.

B.5.3 INTERNAL-STATE ESTIMATORS

Internal-state estimators use latent computations of the model – attention weights or hidden representations – rather than only the response distribution. Their common premise is that predictive unreliability may already be visible in the model's internal processing dynamics before it is fully reflected in the response token distribution.

AttentionScore. AttentionScore summarizes diagonal self-attention weights in the scored input–response sequence. Low diagonal weights increase the score; their association with response errors is evaluated empirically. Writing $\alpha_{\ell,h}(t,t) \in [0,1]$ for the diagonal self-attention weight at layer ℓ, head h, and position t, the executed score is

$$U_{\text{Attn}}(x; \hat{y}) \;:=\; -\frac{1}{H} \sum_{h=1}^{H} \sum_{t \in \mathcal{I}(x,\hat{y})} \log\big(\alpha_{\ell,h}(t,t) + \varepsilon\big),$$

where $\mathcal{I}(x, \hat{y})$ denotes the token positions in the forward-pass attention matrix. The executed setting uses `gen_only=False`, so input positions also contribute; this is not a response-only score. Here $\varepsilon = 10^{-12}$ is a numerical stabilizer and the layer ℓ is the model's middle layer ($\ell = L/2$), as selected by LM-POLYGRAPH when no layer is explicitly specified. This choice matters: attention-based diagnostics vary substantially across layers, and the score should be understood as a layer-specific instantiation of the general idea (Sriramanan et al., 2024).

RAUQ (Recurrent Attention Uncertainty Quantification). RAUQ propagates a confidence signal recursively along the sequence, combining the local token probability with attention from the current token to the previous one, and aggregates the per-layer scores only over the middle third of layers. Let $\mathcal{L} = \{\lfloor L/3 \rfloor, \ldots, \lceil 2L/3 \rceil\}$ denote the indices of the middle third of the model's L attention layers. For each $\ell \in \mathcal{L}$, the head attending most strongly (on average) to the previous token is

$$h^*(\ell) = \arg\max_h \frac{1}{T-1} \sum_{t=2}^{T} \alpha_{\ell,h}(t, t-1).$$

The recurrent confidence at layer ℓ is initialized at $C_1^{(\ell)} = p_{\boldsymbol{\theta}}(\hat{y}_1 \mid x)$ and updated for $t > 1$ by

$$C_t^{(\ell)} = \alpha \cdot p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t}) + (1-\alpha) \cdot \alpha_{\ell,h^*(\ell)}(t, t-1) \cdot C_{t-1}^{(\ell)},$$

where $\alpha \in [0,1]$ is a mixing weight. The per-layer score is $1 - \frac{1}{T}\sum_t \log C_t^{(\ell)}$, and the estimator returns the worst (largest) layer over the middle third:

$$U_{\text{RAUQ}}(x; \hat{y}) \;:=\; \max_{\ell \in \mathcal{L}} \left[1 - \frac{1}{T} \sum_{t=1}^{T} \log C_t^{(\ell)} \right].$$

Unlike AttentionScore, which evaluates each position independently, RAUQ captures how confidence accumulates or degrades along the sequence (Vazhentsev et al., 2025). When α is not specified, LM-POLYGRAPH selects it through two binary flags: $\mathbf{1}_{\text{ent}} = 1$ if the entropy-based variant is used (and 0 for the probability variant), and $\mathbf{1}_{\text{instr}} = 1$ if the underlying model is instruction-tuned (and 0 otherwise). The default is

$$\alpha \;=\; (1 - \mathbf{1}_{\text{ent}})\,(0.2 + 0.3 \cdot \mathbf{1}_{\text{instr}}) \;+\; \mathbf{1}_{\text{ent}}\,(0.8 + 0.1 \cdot \mathbf{1}_{\text{instr}}),$$

which yields $\alpha \in \{0.2, 0.5, 0.8, 0.9\}$ depending on the flags. We use the probability variant ($\mathbf{1}_{\text{ent}} = 0$) on instruction-tuned models ($\mathbf{1}_{\text{instr}} = 1$), so the executed $\alpha = 0.5$.

Contextualized Sequence Likelihood (CSL). Contextualized Sequence Likelihood reweights token negative log-likelihoods by an attention-derived saliency, so that positions the model deems important when producing the final response contribute more to the score than filler positions. Let $\bar{\alpha}_t = \text{mean}_{\ell,h}\, \alpha_{\ell,h}(\text{last}, t)$ be the mean attention paid to position t from the final generation step, averaged over layers and heads, and let $w_t = \bar{\alpha}_t / \sum_{t'} \bar{\alpha}_{t'}$ be the corresponding saliency weight ($w_t \geq 0$, $\sum_t w_t = 1$). The executed score is

$$U_{\text{CSL}}(x; \hat{y}) \;:=\; \sum_{t=1}^{T} w_t \cdot \big[-\log p_{\boldsymbol{\theta}}(\hat{y}_t \mid x, \hat{y}_{<t})\big].$$

The score has the same form as TokenSAR but uses an internal saliency signal in place of an external cross-encoder relevance signal (Lin et al., 2024a).

EigenScore. EigenScore measures the geometric spread of sampled internal embeddings in representation space. Let $\boldsymbol{e}^{(s)} \in \mathbb{R}^d$ denote the decoder hidden state of the last generated token for sample s, and let $\boldsymbol{E} = [\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(S)}] \in \mathbb{R}^{d \times S}$ be the matrix whose columns are the S sample embeddings. Writing $\boldsymbol{J}_d = \boldsymbol{I}_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top \in \mathbb{R}^{d \times d}$ for the centering matrix and α for a regularization scalar, the regularized centered covariance matrix is

$$\boldsymbol{C} = \boldsymbol{E}^\top \boldsymbol{J}_d \boldsymbol{E} + \alpha \boldsymbol{I}_S \;\in\; \mathbb{R}^{S \times S},$$

with eigenvalues $\lambda_1, \ldots, \lambda_S$. The executed score is the mean log-eigenvalue,

$$U_{\text{Eigen}}(x) \;:=\; \frac{1}{S} \sum_{k=1}^{S} \log \lambda_k.$$

Sampled responses whose embeddings cluster in a low-dimensional subspace yield small eigenvalues and a low score (low uncertainty); sampled responses whose embeddings span many directions yield large eigenvalues and a high score (high uncertainty) (Chen et al., 2023). We use $\alpha = 10^{-3}$.

B.5.4 TRAINING-BASED ESTIMATORS

Training-based estimators fit a reference distribution in representation space on a held-out training partition and score test inputs by their distance from that reference. They are best read as familiarity or out-of-distribution signals rather than as readouts of response confidence. All variants below operate on decoder hidden states averaged over the generated tokens of $\hat{y}$, and are all fitted to each task's own training split. The relative variants additionally subtract a background distance term fitted on a broad, general-purpose corpus that we describe below.

Mahalanobis Distance. Mahalanobis Distance is the basic representation-space estimator. Let $\boldsymbol{e}(x, \hat{y}) \in \mathbb{R}^d$ denote the decoder hidden state averaged over all generated tokens of $\hat{y}$, and let $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ be the mean and covariance estimated from the task's training partition. The executed score is the Mahalanobis distance between $\boldsymbol{e}(x, \hat{y})$ and the training centroid:

$$U_{\text{MD}}(x; \hat{y}) \;:=\; \sqrt{\big(\boldsymbol{e}(x, \hat{y}) - \boldsymbol{\mu}\big)^\top \boldsymbol{\Sigma}^{-1} \big(\boldsymbol{e}(x, \hat{y}) - \boldsymbol{\mu}\big)}.$$

A test input whose representation lies near the training centroid receives a small distance (low uncertainty); an input far from the training distribution receives a large distance (high uncertainty) (Lee et al., 2018).

Relative Mahalanobis Distance. Relative Mahalanobis Distance subtracts a background Mahalanobis distance, fitted on a broad general-purpose corpus, in order to isolate task-specific novelty from generic rarity:

$$U_{\mathrm{RMD}}(x;\hat{y}) \;:=\; U_{\mathrm{MD}}(x;\hat{y}) - U_{\mathrm{MD}_0}(x;\hat{y}),$$

where U_{MD_0} uses the background mean and covariance in place of the task mean and covariance. A test input far from both the task and the background receives a small score (general rarity); an input far from the task but close to the background receives a large score, indicating domain shift specific to the task (Ren et al., 2021). We fit the background distribution on `allenai/c4` (shard `en/c4-train.00000-of-01024.json.gz`).[5]

Robust Density Estimation (RDE). Robust Density Estimation replaces the Gaussian fit of Mahalanobis Distance with a robust covariance estimate computed over a Kernel-PCA-projected representation space, to better handle non-elliptical embedding distributions and reduce sensitivity to training-set outliers. Let $\tilde{\boldsymbol{e}}(x;\hat{y}) \in \mathbb{R}^{d'}$ denote the Kernel-PCA projection of $\boldsymbol{e}(x,\hat{y})$, and let $\hat{\boldsymbol{\mu}} \in \mathbb{R}^{d'}$ and $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{d'\times d'}$ be the mean and covariance estimated by the Minimum Covariance Determinant procedure in the projected space. The executed score is the Mahalanobis distance in this robust, dimensionality-reduced space:

$$U_{\mathrm{RDE}}(x;\hat{y}) \;:=\; \sqrt{\big(\tilde{\boldsymbol{e}}(x,\hat{y}) - \hat{\boldsymbol{\mu}}\big)^{\top} \hat{\boldsymbol{\Sigma}}^{-1} \big(\tilde{\boldsymbol{e}}(x,\hat{y}) - \hat{\boldsymbol{\mu}}\big)}.$$

(Yoo et al., 2022).

HUQ-MD and HUQ-RMD. The HUQ variants combine a representation-space distance (epistemic-novelty signal) with a probability-based uncertainty (aleatoric signal) via *empirical-CDF rank fusion* (Vazhentsev et al., 2023). Because Perplexity and Mahalanobis Distance have different scales and distributions, one cannot combine them by raw averaging; rank fusion instead converts each score to a scale-free percentile relative to the training partition before averaging. Concretely, let $\mathcal{D}$ denote the task's training partition. The empirical CDF of U_{PPL} on that partition is

$$F^{\mathrm{train}}_{\mathrm{PPL}}(u) = \frac{1}{|\mathcal{D}|} \sum_{x' \in \mathcal{D}} \mathbf{1}\big[U_{\mathrm{PPL}}(x', \hat{y}(x')) \le u\big] \;\in\; [0,1],$$

and $F^{\mathrm{train}}_{\mathrm{MD}}$ is defined analogously for U_{MD}. Evaluating these CDFs at a test input gives its percentile rank on each score (the fraction of training inputs with a lower score), and the executed HUQ-MD score is the mean of the two percentile ranks:

$$U_{\text{HUQ-MD}}(x;\hat{y}) \;:=\; \tfrac{1}{2}\Big[F^{\mathrm{train}}_{\mathrm{PPL}}\big(U_{\mathrm{PPL}}(x;\hat{y})\big) + F^{\mathrm{train}}_{\mathrm{MD}}\big(U_{\mathrm{MD}}(x;\hat{y})\big)\Big] \;\in\; [0,1].$$

HUQ-RMD is defined analogously with U_{RMD} in place of U_{MD}. Because both components are converted to $[0,1]$-valued percentile ranks before averaging, the combination is scale-free and distribution-free.

B.5.5 Reflexive estimators (white-box)

Reflexive estimators prompt the model to judge its own answer. The white-box variants additionally require direct access to the probability assigned to the "True" option; their black-box counterpart is described in Section B.6.4.

PTrue. Given an input x and a proposed answer $\hat{y}$, PTrue constructs a self-evaluation prompt $\mathrm{prompt}_{\mathrm{PT}}(x;\hat{y})$ asking the model to judge whether $\hat{y}$ is correct and reads off the probability assigned to the "True" option. The executed score is

$$U_{\mathrm{PTrue}}(x;\hat{y}) \;:=\; -\log p_{\boldsymbol{\theta}}\big(\text{"True"} \mid \mathrm{prompt}_{\mathrm{PT}}(x;\hat{y})\big).$$

A high probability of "True" yields a small score (low uncertainty); a low probability yields a large score (high uncertainty) (Kadavath et al., 2022). The exact prompt template is:

```
Question: {x}.
Proposed Answer: {ŷ}.
Is the proposed answer True or False?
```

[5] https://huggingface.co/datasets/allenai/c4

PTrue Sampling. PTrue Sampling augments the self-evaluation prompt with S sampled candidate answers $y^{(1)}, \ldots, y^{(S)}$ before asking the truth judgment, in order to expose the model to its own response variability:

$$U_{\text{PTrue-S}}(x; \hat{y}) \;:=\; -\log p_{\boldsymbol{\theta}}\big(\text{“True”} \mid \text{prompt}_{\text{PT-S}}(x, \hat{y}, y^{(1)}, \ldots, y^{(S)})\big).$$

The exact prompt template is:

```
Question: {x}
Here are some ideas that were brainstormed: {y^(1),...,y^(S)}
Possible answer: {ŷ}
Is the possible answer:
(A) True
(B) False
The possible answer is:
```

B.6 Black-box estimators

Black-box estimators operate only on sampled responses and external semantic/lexical-relation functions, without using token probabilities, attention, or hidden states. We organize them into four groups: the simplest estimators that summarize the semantic-class structure of the samples (NumSet, LabelProb); spectral and graph-density estimators on a pairwise similarity graph (Kernel Language Entropy, EigValLaplacian, Eccentricity, DegMat, LUQ); lexical baselines that replace semantic relations with pairwise surface overlap (ROUGE-L Similarity, BLEU Similarity); and a reflexive black-box estimator (PTrue Empirical).

B.6.1 Semantic-class summaries

NumSet. NumSet is the simplest semantic-consistency estimator: it counts the number of distinct semantic classes among the sampled responses. Samples are partitioned into bidirectional-entailment classes $C_1, \ldots, C_M$, as in Semantic Entropy, and the executed score is the cluster count:

$$U_{\text{NumSet}}(x) \;:=\; M.$$

The score is integer-valued, with a minimum of 1 when all samples express the same meaning and a maximum of S when every sample expresses a different meaning (Lin et al., 2024b). The granularity is coarse, but the score is robust to NLI-model calibration.

LabelProb. We retain the package name LabelProb for consistency with LM-Polygraph, but the executed implementation is more accurately described as the complement of the dominant semantic-class frequency:

$$U_{\text{LabelProb}}(x) \;:=\; 1 - \max_{1 \le m \le M} \frac{|C_m|}{S}.$$

The score is zero when a single semantic class accounts for all samples, and approaches $1 - 1/S$ when every sample belongs to a different class. LabelProb is therefore best understood as a majority-class complement, not as a black-box approximation of maximum sequence probability as the name might suggest (Vashurin et al., 2025a).

B.6.2 Spectral and graph-density estimators

These estimators build a pairwise relation matrix $\boldsymbol{W} = (W_{ij})_{i,j=1}^{S} \in \mathbb{R}^{S \times S}$ over sampled responses and summarize its topology. Three of them rely on spectral analysis of the corresponding graph (Kernel Language Entropy uses kernel eigenvalues; EigValLaplacian uses Laplacian eigenvalues; Eccentricity uses Laplacian eigenvectors), and the remaining two summarize the graph through its degree structure (DegMat) or through average pairwise scores (LUQ). For EigValLaplacian, DegMat, and Eccentricity, we run three edge constructions: NLI-entailment ($W_{ij} = E_{ij}$), NLI-contradiction ($W_{ij} = 1 - C_{ij}$), and Jaccard token similarity.

Kernel Language Entropy (KLE). KLE quantifies semantic uncertainty as the von Neumann entropy of a heat-diffusion kernel on the semantic graph. Let $\boldsymbol{D} = \mathrm{diag}\big(\sum_j W_{ij}\big)$ be the degree matrix, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ the unnormalized graph Laplacian, and $\boldsymbol{K} = \exp(-t\boldsymbol{L})$ the heat kernel (subsequently row-normalized to unit trace). Writing $\lambda_k(\boldsymbol{K})$ for the eigenvalues of $\boldsymbol{K}$:

$$U_{\mathrm{KLE}}(x) \;:=\; -\operatorname{Tr}\big(\boldsymbol{K}\log\boldsymbol{K}\big) \;=\; -\sum_k \lambda_k(\boldsymbol{K})\log\lambda_k(\boldsymbol{K}).$$

A graph with many disconnected semantic clusters yields several large eigenvalues and a high entropy; a fully connected graph yields a single dominant eigenvalue and low entropy (Nikitin et al., 2024).

EigValLaplacian. EigValLaplacian also operates on the spectrum of the graph, but on the eigenvalues of the normalized Laplacian $\tilde{\boldsymbol{L}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ rather than on a heat kernel. Writing $\lambda_1, \ldots, \lambda_S \in [0, 2]$ for its eigenvalues:

$$U_{\mathrm{EVL}}(x) \;:=\; \sum_{k=1}^{S} \max(0, 1 - \lambda_k).$$

Each term measures how close one mode of variation is to being fully disconnected; the sum is a continuous relaxation of the number of semantic sets (Lin et al., 2024b).

Eccentricity. Eccentricity (Ecc) uses the eigen*vectors* of $\tilde{\boldsymbol{L}}$ rather than its eigenvalues. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ be the eigenvectors of $\tilde{\boldsymbol{L}}$ corresponding to its smallest k eigenvalues, and stack their per-sample entries into spectral embeddings $\boldsymbol{v}_i = \big(\boldsymbol{u}_1[i], \ldots, \boldsymbol{u}_k[i]\big) \in \mathbb{R}^k$ for sample i. Writing $\bar{\boldsymbol{v}} = \frac{1}{S}\sum_i \boldsymbol{v}_i$ for the centroid:

$$U_{\mathrm{Ecc}}(x) \;:=\; \left(\sum_{i=1}^{S} \|\boldsymbol{v}_i - \bar{\boldsymbol{v}}\|_2^2\right)^{1/2}.$$

A graph in which all samples occupy a tight region of the spectral embedding space yields a small score; a graph with spectral outliers yields a large score (Lin et al., 2024b).

DegMat. Degree Matrix (DegMat) summarizes a graph's degree structure rather than its spectrum. The executed score is the mean missing pairwise similarity:

$$U_{\mathrm{DegMat}}(x) \;:=\; \frac{S^2 - \sum_{i=1}^{S}\sum_{j=1}^{S} W_{ij}}{S^2},$$

which equals zero when all pairs have $W_{ij} = 1$ (full mutual agreement) and approaches 1 when no pairs agree. Like Eccentricity, DegMat depends on the choice of edge construction, and we report all three variants (Lin et al., 2024b).

LUQ. Long-text Uncertainty Quantification (LUQ) summarizes the graph through average pairwise NLI confidence, using the continuous softmax score from the NLI logits rather than discretized entail/contradict decisions. Writing $\sigma_{ij} = \exp(z_{ij}^E)/\big(\exp(z_{ij}^E) + \exp(z_{ij}^C)\big)$ for the soft entailment confidence between $y^{(i)}$ and $y^{(j)}$:

$$U_{\mathrm{LUQ}}(x) \;:=\; 1 - \frac{1}{S(S-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{S} \sigma_{ij}.$$

A high average pairwise entailment confidence yields a small score (low uncertainty); weak average confidence yields a large score (Zhang et al., 2024).

B.6.3 Lexical similarity baselines

ROUGE-L Similarity and BLEU Similarity. These baselines replace NLI-based semantic relations with pairwise surface-overlap measures. For $\mathrm{metric}(\cdot,\cdot) \in \{\mathrm{ROUGE\text{-}L}, \mathrm{BLEU}\}$, the executed score is the negated mean pairwise similarity:

$$U_{\mathrm{LexSim}}(x) \;:=\; -\frac{2}{S(S-1)} \sum_{1 \le i < j \le S} \mathrm{metric}(y^{(i)}, y^{(j)}).$$

Table 2: Numerical hyperparameters and external models for every estimator executed in our benchmark. "Samples" is the number of stochastic generations generated per input. Estimators that have no tunable hyperparameters and no external models (MSP, Perplexity, Mean Token Entropy, SelfCertainty, PMI, MC-SE, MC-NSE, CSL) are omitted from the table. Across all sample-based estimators, the sample pool is shared, the sampling temperature is 1, and the random seed is 42.

Estimator	Samples	Hyperparameters and external models
logit-based (Sec. B.5.1)		
Rényi Divergence	1	$\alpha = 0.5, \tau = 2$
Fisher–Rao Distance	1	$\tau = 2$
CPMI	1	$\tau = 0.0656, \lambda = 3.599$
TokenSAR	1	cross-encoder `stsb-roberta-large`
CCP	1	NLI `deberta-large-mnli`; top-k alternatives, $k = 10$
Sample semantic dispersion (Sec. B.5.2)		
MC-SE, MC-NSE	$S = 10$	sampling temperature 1
Semantic Entropy	$S = 10$	NLI `deberta-large-mnli`
Semantic Density	$S = 10$	NLI `deberta-large-mnli`
SentenceSAR	$S = 10$	cross-encoder `stsb-roberta-large`; $\tau = 1$
SAR	$S = 10$	cross-encoder `stsb-roberta-large`; $\tau = 1$
Cocoa{MSP, PPL, MTE}	$S = 10$	cross-encoder `stsb-roberta-large`;
Internal state (Sec. B.5.3)		
AttentionScore	1	layer $\ell = L/2$; $\varepsilon = 10^{-12}$
RAUQ	1	$\alpha = 0.5$; middle-third layers $\mathcal{L} = \{\lfloor L/3 \rfloor, \ldots, \lceil 2L/3 \rceil\}$
EigenScore	$S = 10$	decoder hidden state of last token; $\alpha = 10^{-3}$
Training-based (Sec. B.5.4)		
Mahalanobis Distance	1	decoder hidden states; fitted per task on training partition
Relative MD	1	background: `allenai/c4`, shard `en/c4-train.00000`
RDE	1	Kernel-PCA projection + MCD covariance
HUQ-MD, HUQ-RMD	1	CDF rank fusion of U_{PPL} and U_{MD} / U_{RMD} on training partition
Reflexive, white-box (Sec. B.5.5)		
PTrue	1	prompt: see §B.5.5
PTrue Sampling	$S = 10$	prompt: see §B.5.5
Black-box (Sec. B.6)		
NumSet, LabelProb	$S = 10$	NLI `deberta-large-mnli`
KLE	$S = 10$	NLI `deberta-large-mnli`; heat-kernel time $t = 0.3$
EigValLaplacian, Ecc, DegMat	$S = 10$	NLI `deberta-large-mnli`;
LUQ	$S = 10$	NLI `deberta-large-mnli`
Lexical Similarity	$S = 10$	metrics: ROUGE-L, BLEU
PTrue Empirical	$N = 10$ self-eval.	prompt: see §B.6.4

Surface-form variation across samples yields a high score even when the samples convey the same meaning, so these baselines cannot distinguish valid paraphrasing from semantic disagreement (Lin, 2004; Papineni et al., 2002). We report ROUGE-L and BLEU variants separately.

B.6.4 Reflexive estimator (black-box)

PTrue Empirical. PTrue Empirical is the black-box counterpart of PTrue: instead of reading off the probability of the "True" token, it draws N stochastic self-evaluation responses from the model and measures the empirical frequency with which the model judges its own answer to be true:

$$U_{\text{PTrue-E}}(x; \hat{y}) \;:=\; 1 - \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\Big[\text{"True"} \in \text{sample}_n\big(\text{prompt}_{\text{PT}}(x; \hat{y})\big)\Big],$$

using the same self-evaluation prompt as PTrue. We use $N = 10$ (Tian et al., 2023).

C Additional methodology

These details complete the benchmark setup and evaluation methodology in Sections 3 and 4. Procedures for the sensitivity analyses appear beside their results in Section D.

C.1 Datasets and response evaluation

This subsection completes the dataset descriptions in Section 3.1 with the quality formula, label processing and sample counts.

For LongWiki, let m be the number of extracted response claims and s the number judged supported. Precision is $P = s/m$, Recall@K is $R_K = \min(s/K, 1)$, and their harmonic mean is $F_K = 2PR_K/(P + R_K)$, with $K = 32$. We assign $F_K = 0$ when both components are zero. The recall component measures supported claims relative to a fixed quota, not coverage of an independently enumerated reference answer.

We use the `Meta-Llama-3.1-70B-Instruct-AWQ-INT4` checkpoint for the primary HalluLens evaluation. PreciseWikiQA's correctness prompt requests correct, incorrect, or unverifiable judgments, but also contains a Yes/No instruction. The parser accepts `CORRECT`/`YES` as correct and `INCORRECT`/`NO` as incorrect. An abstention flag can overlap a correctness label. The evaluation retains verifiable judgments without an abstention flag; The tables below report the retained responses and the original judgment categories.

LongWiki responses without a quality target after the knowledge gate are excluded: 4, 1, and 3 of the 2,000 responses for L7B, L13B, and M7B, respectively. Table 4 reports response quality and claim counts.

Table 3: Response counts used in the main evaluation. Excluded responses have no usable target or fail the task-specific inclusion rule. RAGTruth uses human annotation; the other tasks use the Llama judge.

Task	Model	Labels	Evaluated	Errors	Excluded
RAGTruth	Llama 2 7B	Human	1,959	1,207	0
RAGTruth	Llama 2 13B	Human	1,959	1,110	0
RAGTruth	Mistral 7B	Human	1,959	1,308	0
PreciseWikiQA	Llama 2 7B	Llama judge	1,767	1,637	233
PreciseWikiQA	Llama 2 13B	Llama judge	1,923	1,672	77
PreciseWikiQA	Mistral 7B	Llama judge	1,955	1,686	45
LongWiki	Llama 2 7B	Llama judge	1,996	1,480	4
LongWiki	Llama 2 13B	Llama judge	1,999	1,729	1
LongWiki	Mistral 7B	Llama judge	1,997	1,785	3
NonExistentRefusal	Llama 2 7B	Llama judge	2,000	1,928	0
NonExistentRefusal	Llama 2 13B	Llama judge	2,000	1,665	0
NonExistentRefusal	Mistral 7B	Llama judge	2,000	1,875	0

Table 4: LongWiki quality and claim counts among responses with a valid target. Precision is the fraction of supported response claims; Recall@32 is the number of supported claims divided by 32, capped at one. F1@32 is their harmonic mean. Retrieved passages are counted once per response; these are passage counts, not counts of reference claims.

Quantity	Llama 2 7B	Llama 2 13B	Mistral 7B
Responses with F1 target	1,996	1,999	1,997
Mean precision	0.3671	0.4176	0.2679
Mean Recall@32	0.4056	0.2519	0.3258
Mean F1@32	0.3796	0.3104	0.2906
Response claims: mean	35.48	19.24	38.93
Response claims: median	36	19	39
Response claims: min–max	2–64	1–37	8–67
Retrieved passages: mean	17.32	14.65	19.32
Retrieved passages: median	16	14	18
Retrieved passages: min–max	3–69	3–49	3–81

Table 5: PreciseWikiQA judgment counts and percentages among all 2,000 responses per model. Abstention overlaps correctness: one Llama 2 7B response is both correct and abstaining. The parser treats `No` as incorrect. Evaluated response counts, after excluding abstentions and unverifiable judgments, are in Table 3.

Judgment	Llama 2 7B	Llama 2 13B	Mistral 7B
Correct	131 (6.55%)	251 (12.55%)	269 (13.45%)
Incorrect	1,868 (93.40%)	1,747 (87.35%)	1,730 (86.50%)
Unverifiable	0 (0.00%)	1 (0.05%)	1 (0.05%)
Incorrect alias: `No`	1 (0.05%)	1 (0.05%)	0 (0.00%)
Abstaining	233 (11.65%)	76 (3.80%)	44 (2.20%)

C.2 AlignScore reference texts

The reference-based comparator in Section 3.2 requires a reference text and a scoring direction, specified here.

AlignScore (Zha et al., 2023) estimates whether one text is supported by another. For PreciseWikiQA and LongWiki, $\mathrm{AlignScore}(y, \mathrm{ref})$ checks the response against a gold answer (AS-Precision-gold) or source Wikipedia page (AS-Precision-Wikipedia), measuring a precision score with respect to the reference. Reversing the arguments, $\mathrm{AlignScore}(\mathrm{ref}, y)$ (AS-Recall), measures how much of the reference the response recovers, giving a recall-like score. For NonExistentRefusal, no gold answer is available, so the reference is the abstention template below. The complete scores appear in Section D.1.

Abstention reference. This text is used as an AlignScore reference for refusal responses. It is not the prompt used by the judge to assign abstention labels.

> **Abstention template.** "No information is available about this, and I am not aware of any such entity being recognized or documented. This name does not appear to correspond to anything known, and I do not believe this exists as a real or established thing. There may be some confusion, or this could be a misspelling or an alternative rendering of a different name. I am not certain what this refers to, and this term is not widely recognized in any domain I can identify. If you could provide more context or clarify your intent, that would help determine whether a related concept exists."

C.3 Evaluation metrics

The following conventions specify the metrics introduced in Section 4.

We compute AUPRC as average precision. Trapezoidal integration can overestimate the precision–recall area (Davis & Goadrich, 2006); for a constant score, it gives half of one plus the error rate, rather than the error rate itself. RCE uses the plug-in ℓ_1 estimator with ten equal-mass bins and the native quality target, including continuous F1@32 for LongWiki.

For PRR, responses are ordered from lowest to highest uncertainty, preserving input order within score ties. At rejection rate r, the retained count is $n - \lfloor rn \rfloor$, and retained quality is the mean quality of those responses. We integrate 101 points over rejection rates from 0 to 1. The empty-set endpoint is assigned quality one. The oracle and random curves use the same grid; the random baseline averages 1,000 permutations with seed 42. Negative PRR values are retained.

For metrics with bootstrap estimates, we use 1,000 resamples and report their standard deviation. Sensitivity results without bootstrap estimates are descriptive point estimates.

D Additional results

This appendix gives the complete results behind Section 5, followed by the target and control checks.

D.1 Complete performance tables

The tables report each task and model separately. AlignScore uses reference information and is grouped separately from uncertainty estimators. P(True) Sampling was not run on LongWiki or NonExistentRefusal because of resource constraints; the other 45 estimators have complete coverage.

Table 6: AUROC ↑, PRR ↑, RCE ↓ for **All Models** across tasks (higher AUROC/PRR and lower RCE are better; color shows numerical value). Uncertainty estimators and AlignScore variants are ranked separately: bold = best per column (ties marked); underline = second-best (ties marked) within the uncertainty group; AlignScore group shows only the best.

	Estimator	RAGTruth			PreciseWikiQA			LongWiki			NonExistentRefusal		
		AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓
Baseline	Output Token Count	0.69±0.06	**0.40±0.10**	**0.10±0.05**	0.61±0.02	0.15±0.07	0.16±0.02	0.49±0.01	−0.05±0.04	0.36±0.02	0.70±0.12	0.41±0.24	0.25±0.02
Information-based	Maximum Seq. Probability (MSP)	0.64±0.10	0.33±0.17	0.16±0.10	0.71±0.05	0.36±0.11	0.13±0.03	0.56±0.03	0.06±0.05	0.19±0.06	0.67±0.08	0.31±0.18	0.18±0.08
	Perplexity (PPL)	0.44±0.09	−0.03±0.18	0.40±0.12	0.64±0.07	0.20±0.11	0.16±0.08	0.57±0.03	0.10±0.05	0.16±0.05	0.43±0.13	−0.06±0.14	0.35±0.11
	Mean Token Entropy (MTE)	0.52±0.01	0.17±0.02	0.30±0.03	0.63±0.07	0.18±0.13	0.18±0.07	0.57±0.03	0.10±0.05	0.16±0.05	0.44±0.13	−0.06±0.13	0.35±0.11
	Pointwise Mutual Info. (PMI)	0.33±0.01	−0.28±0.08	0.48±0.02	0.46±0.01	−0.02±0.02	0.41±0.01	0.51±0.02	0.02±0.02	0.31±0.03	0.44±0.02	−0.09±0.06	0.35±0.03
	Conditional PMI	0.51±0.07	0.14±0.09	0.30±0.09	0.31±0.03	−0.22±0.03	0.48±0.03	0.43±0.03	−0.15±0.06	0.47±0.03	0.57±0.13	0.17±0.22	0.29±0.14
	Self Certainty	0.46±0.02	0.05±0.03	0.37±0.04	0.56±0.02	0.03±0.03	0.23±0.04	0.55±0.02	0.07±0.03	0.19±0.04	0.50±0.08	0.02±0.09	0.28±0.06
	Rényi Divergence	0.45±0.02	0.02±0.02	0.40±0.03	0.54±0.01	0.02±0.03	0.26±0.03	0.54±0.02	0.06±0.02	0.21±0.05	0.50±0.07	0.03±0.08	0.27±0.05
	Fisher–Rao Distance	0.47±0.03	0.07±0.04	0.36±0.05	0.53±0.02	0.01±0.04	0.26±0.05	0.55±0.02	0.07±0.03	0.18±0.05	0.53±0.07	0.08±0.07	0.25±0.03
	Token SAR	0.44±0.09	−0.03±0.18	0.40±0.12	0.70±0.05	0.32±0.09	0.12±0.04	0.57±0.03	0.11±0.05	0.15±0.05	0.42±0.15	−0.06±0.15	0.35±0.10
	Claim-Conditioned Probability (CCP)	0.61±0.11	0.26±0.23	0.19±0.13	0.72±0.02	0.35±0.08	0.10±0.02	0.62±0.02	0.19±0.08	0.10±0.04	**0.73±0.09**	0.41±0.22	**0.15±0.05**
Sampling-based	Monte-Carlo Seq. Entropy	0.42±0.06	−0.04±0.12	0.40±0.05	0.67±0.07	0.29±0.11	0.14±0.05	0.52±0.01	0.03±0.03	0.27±0.04	0.58±0.06	0.13±0.09	0.24±0.06
	Monte-Carlo Norm. Seq. Entropy	0.37±0.02	−0.17±0.02	0.48±0.01	0.66±0.06	0.26±0.10	0.14±0.06	0.53±0.01	0.06±0.03	0.23±0.06	0.48±0.08	−0.01±0.12	0.36±0.12
	Semantic Entropy	0.42±0.07	−0.05±0.14	0.41±0.06	0.68±0.07	0.31±0.11	0.14±0.03	0.54±0.03	0.06±0.05	0.23±0.06	0.58±0.06	0.11±0.10	0.25±0.07
	Semantic Density	0.45±0.05	−0.03±0.07	0.40±0.04	**0.75±0.03**	**0.45±0.07**	0.11±0.05	0.59±0.02	0.17±0.06	0.11±0.03	0.42±0.07	−0.08±0.08	0.42±0.06
	Sentence SAR	0.51±0.08	0.12±0.15	0.31±0.08	0.71±0.06	0.38±0.12	0.14±0.03	0.51±0.02	0.00±0.03	0.31±0.07	0.66±0.05	0.21±0.10	0.15±0.01
	SAR	0.41±0.05	−0.08±0.08	0.43±0.05	0.71±0.05	0.37±0.09	0.12±0.02	0.59±0.02	0.20±0.09	0.13±0.03	0.43±0.08	−0.09±0.09	0.40±0.09
	Cocoa MSP	0.62±0.10	0.28±0.18	0.19±0.12	0.73±0.04	0.40±0.09	0.12±0.01	0.61±0.03	0.18±0.08	0.13±0.06	0.60±0.08	0.18±0.14	0.22±0.10
	Cocoa PPL	0.45±0.07	0.00±0.15	0.39±0.11	0.68±0.06	0.29±0.11	0.14±0.05	0.62±0.04	0.22±0.09	0.11±0.04	0.39±0.12	−0.11±0.12	0.41±0.07
	Cocoa MTE	0.51±0.02	0.12±0.04	0.31±0.04	0.68±0.06	0.28±0.13	0.12±0.03	0.62±0.04	0.22±0.09	0.10±0.05	0.40±0.12	−0.11±0.12	0.40±0.08
Internal	Attention Score	**0.69±0.03**	0.39±0.07	0.13±0.01	0.59±0.03	0.11±0.05	0.18±0.03	0.47±0.01	−0.09±0.03	0.45±0.03	0.72±0.11	**0.44±0.24**	0.20±0.02
	RAUQ	0.44±0.09	−0.02±0.18	0.39±0.12	0.68±0.09	0.33±0.13	0.17±0.09	0.56±0.03	0.09±0.05	0.16±0.05	0.44±0.12	−0.05±0.12	0.35±0.11
	CSL	0.39±0.02	−0.13±0.08	0.46±0.04	0.62±0.05	0.20±0.10	0.17±0.05	0.53±0.03	0.05±0.03	0.23±0.04	0.47±0.01	0.04±0.10	0.38±0.01
	EigenScore	0.44±0.08	−0.03±0.16	0.40±0.06	0.61±0.06	0.23±0.12	0.24±0.04	0.47±0.03	−0.07±0.06	0.41±0.07	0.56±0.10	0.10±0.18	0.31±0.11
Training-based	Mahalanobis Distance (MD)	0.27±0.04	−0.48±0.09	0.48±0.00	0.50±0.02	−0.02±0.05	0.32±0.04	0.40±0.04	−0.13±0.08	0.45±0.05	0.35±0.18	−0.05±0.23	0.40±0.14
	Relative MD (RMD)	0.29±0.03	−0.41±0.05	0.49±0.00	0.48±0.02	−0.05±0.04	0.35±0.03	0.42±0.05	−0.10±0.08	0.43±0.08	0.30±0.15	−0.13±0.14	0.43±0.10
	Robust Density Estimation (RDE)	0.32±0.05	−0.41±0.11	0.46±0.02	0.51±0.09	0.01±0.10	0.29±0.14	0.49±0.06	−0.04±0.09	0.36±0.14	0.36±0.12	−0.10±0.13	0.44±0.09
	HUQ-MD	0.44±0.09	−0.03±0.18	0.40±0.12	0.64±0.07	0.19±0.10	0.16±0.09	0.57±0.03	0.10±0.05	0.16±0.05	0.43±0.13	−0.06±0.14	0.35±0.11
	HUQ-RMD	0.44±0.08	−0.03±0.17	0.40±0.12	0.64±0.07	0.19±0.10	0.16±0.09	0.57±0.03	0.10±0.05	0.16±0.05	0.43±0.13	−0.06±0.14	0.35±0.11
Reflexive	P(True)	0.51±0.09	0.06±0.19	0.29±0.15	0.50±0.04	−0.03±0.05	0.33±0.07	0.47±0.01	−0.03±0.01	0.38±0.02	0.42±0.15	−0.07±0.10	0.35±0.08
	P(True) Sampling	0.50±0.10	0.02±0.19	0.30±0.16	0.51±0.03	0.01±0.04	0.29±0.06	–	–	–	–	–	–
	P(True) Empirical	0.57±0.08	0.14±0.14	0.21±0.10	0.50±0.02	0.01±0.03	0.32±0.03	0.49±0.00	−0.01±0.01	0.37±0.02	0.45±0.09	−0.07±0.10	0.37±0.09
Black-box	NumSet	0.48±0.01	−0.03±0.01	0.37±0.01	0.73±0.04	0.37±0.07	**0.09±0.02**	0.53±0.01	0.04±0.02	0.26±0.02	0.45±0.01	−0.05±0.01	0.42±0.02
	LabelProb	0.45±0.10	−0.08±0.18	0.35±0.11	0.71±0.04	0.37±0.08	0.12±0.02	0.54±0.03	0.11±0.06	0.20±0.08	0.51±0.06	0.00±0.08	0.33±0.10
	Kernel Language Entropy (KLE)	0.42±0.09	−0.12±0.18	0.40±0.07	0.74±0.04	0.42±0.08	0.12±0.05	0.60±0.03	0.19±0.08	0.11±0.04	0.47±0.05	−0.05±0.09	0.38±0.06
	EigValLap NLI (entail)	0.42±0.09	−0.12±0.18	0.39±0.08	0.74±0.04	0.43±0.10	0.11±0.04	0.59±0.03	0.17±0.07	0.12±0.04	0.47±0.05	−0.06±0.08	0.38±0.07
	EigValLap NLI (contra)	0.41±0.03	−0.21±0.05	0.46±0.02	0.68±0.07	0.33±0.12	0.15±0.07	**0.65±0.03**	**0.29±0.07**	0.08±0.03	0.48±0.04	−0.04±0.05	0.38±0.06
	EigValLap Jaccard	0.36±0.05	−0.16±0.09	0.47±0.03	0.65±0.03	0.24±0.05	0.14±0.04	0.58±0.02	0.17±0.06	0.11±0.04	0.39±0.05	−0.11±0.03	0.47±0.04
	Eccentricity NLI (entail)	0.44±0.11	−0.10±0.20	0.35±0.10	0.72±0.04	0.39±0.09	0.14±0.01	0.58±0.03	0.16±0.07	0.15±0.05	0.50±0.05	−0.02±0.09	0.34±0.07
	Eccentricity NLI (contra)	0.40±0.02	−0.23±0.03	0.46±0.02	0.74±0.04	0.41±0.09	0.10±0.02	0.64±0.02	0.28±0.07	0.08±0.03	0.48±0.05	−0.04±0.05	0.38±0.07
	Eccentricity Jaccard	0.54±0.04	0.11±0.09	0.29±0.07	0.66±0.05	0.28±0.08	0.15±0.04	0.49±0.01	−0.02±0.00	0.39±0.00	0.52±0.01	0.01±0.01	0.29±0.03
	DegMat NLI (entail)	0.43±0.10	−0.11±0.20	0.38±0.09	0.75±0.04	0.43±0.08	0.11±0.05	0.57±0.03	0.15±0.07	0.14±0.06	0.47±0.06	−0.05±0.09	0.36±0.08
	DegMat NLI (contra)	0.38±0.03	−0.27±0.04	0.48±0.01	0.69±0.06	0.33±0.10	0.14±0.06	0.65±0.03	0.29±0.08	**0.07±0.02**	0.48±0.04	−0.04±0.05	0.38±0.06
	DegMat Jaccard	0.38±0.06	−0.11±0.11	0.44±0.04	0.66±0.03	0.27±0.07	0.13±0.04	0.59±0.02	0.18±0.07	0.11±0.04	0.39±0.05	−0.11±0.04	0.46±0.04
	LUQ	0.37±0.04	−0.28±0.06	0.47±0.03	0.74±0.04	0.42±0.08	0.11±0.05	0.65±0.03	0.29±0.08	0.08±0.02	0.45±0.04	−0.07±0.06	0.40±0.05
	Lexical Sim. (ROUGE-L)	0.37±0.06	−0.13±0.11	0.45±0.03	0.67±0.03	0.28±0.06	0.13±0.04	0.55±0.01	0.11±0.03	0.15±0.03	0.33±0.06	−0.17±0.04	0.49±0.01
	Lexical Sim. (BLEU)	0.40±0.06	−0.09±0.12	0.44±0.04	0.65±0.04	0.26±0.07	0.14±0.03	0.57±0.02	0.15±0.06	0.12±0.04	0.30±0.04	−0.19±0.03	0.50±0.01
AlignScore	AS-Precision(gold_answer)	**0.71±0.02**	**0.51±0.03**	**0.09±0.04**	0.82±0.06	0.70±0.07	0.25±0.04	0.65±0.03	0.29±0.06	0.07±0.02	**0.66±0.08**	0.31±0.17	**0.20±0.06**
	AS-Precision(wikipage)	–	–	–	0.81±0.04	0.49±0.14	**0.08±0.02**	**0.67±0.05**	**0.32±0.11**	**0.07±0.04**	–	–	–
	AS-Recall(gold_answer)	0.56±0.04	0.19±0.07	0.23±0.07	**0.90±0.01**	**0.82±0.02**	0.22±0.05	0.59±0.01	0.15±0.04	0.14±0.02	0.65±0.05	**0.39±0.09**	0.29±0.08
	AS-Recall(wikipage)	–	–	–	0.51±0.02	0.02±0.02	0.32±0.03	0.57±0.01	0.10±0.03	0.16±0.03	–	–	–

Table 7: AUROC ↑, PRR ↑, RCE ↓ for **Mistral-7B-Instruct** across tasks (higher AUROC/PRR and lower RCE are better; color shows numerical value). Uncertainty estimators and AlignScore variants are ranked separately: bold = best per column (ties marked); underline = second-best (ties marked) within the uncertainty group; AlignScore group shows only the best.

	Estimator	RAGTruth			PreciseWikiQA			LongWiki			NonExistentRefusal		
		AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓
Baseline	Output Token Count	$0.70_{\pm 0.01}$	$0.45_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.64_{\pm 0.02}$	$0.25_{\pm 0.03}$	$0.14_{\pm 0.04}$	$0.50_{\pm 0.02}$	$-0.04_{\pm 0.02}$	$0.35_{\pm 0.05}$	$0.66_{\pm 0.02}$	$0.35_{\pm 0.05}$	$0.26_{\pm 0.06}$
Information-based	Maximum Seq. Probability (MSP)	$\mathbf{0.76_{\pm 0.01}}$	$0.54_{\pm 0.02}$	$0.06_{\pm 0.02}$	$0.76_{\pm 0.02}$	$0.50_{\pm 0.03}$	$0.09_{\pm 0.03}$	$0.52_{\pm 0.02}$	$0.00_{\pm 0.02}$	$0.26_{\pm 0.05}$	$0.68_{\pm 0.03}$	$0.25_{\pm 0.04}$	$0.16_{\pm 0.04}$
	Perplexity (PPL)	$0.56_{\pm 0.02}$	$0.21_{\pm 0.03}$	$0.23_{\pm 0.03}$	$0.71_{\pm 0.02}$	$0.35_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.53_{\pm 0.02}$	$0.04_{\pm 0.02}$	$0.22_{\pm 0.05}$	$0.55_{\pm 0.03}$	$0.05_{\pm 0.03}$	$0.23_{\pm 0.06}$
	Mean Token Entropy (MTE)	$0.53_{\pm 0.02}$	$0.18_{\pm 0.03}$	$0.28_{\pm 0.03}$	$0.72_{\pm 0.02}$	$0.37_{\pm 0.03}$	$0.09_{\pm 0.02}$	$0.52_{\pm 0.02}$	$0.03_{\pm 0.02}$	$0.23_{\pm 0.05}$	$0.56_{\pm 0.03}$	$0.05_{\pm 0.03}$	$0.23_{\pm 0.06}$
	Pointwise Mutual Info. (PMI)	$0.35_{\pm 0.01}$	$-0.18_{\pm 0.03}$	$0.45_{\pm 0.02}$	$0.46_{\pm 0.02}$	$-0.04_{\pm 0.03}$	$0.43_{\pm 0.05}$	$0.49_{\pm 0.02}$	$0.02_{\pm 0.02}$	$0.31_{\pm 0.05}$	$0.45_{\pm 0.03}$	$-0.03_{\pm 0.03}$	$0.32_{\pm 0.04}$
	Conditional PMI	$0.42_{\pm 0.01}$	$0.01_{\pm 0.03}$	$0.43_{\pm 0.02}$	$0.28_{\pm 0.02}$	$-0.25_{\pm 0.02}$	$0.50_{\pm 0.01}$	$0.47_{\pm 0.02}$	$-0.07_{\pm 0.02}$	$0.43_{\pm 0.04}$	$0.44_{\pm 0.03}$	$-0.05_{\pm 0.03}$	$0.44_{\pm 0.04}$
	Self Certainty	$0.49_{\pm 0.02}$	$0.09_{\pm 0.03}$	$0.31_{\pm 0.04}$	$0.58_{\pm 0.02}$	$0.07_{\pm 0.03}$	$0.18_{\pm 0.04}$	$0.52_{\pm 0.02}$	$0.05_{\pm 0.02}$	$0.23_{\pm 0.06}$	$0.60_{\pm 0.03}$	$0.12_{\pm 0.04}$	$0.21_{\pm 0.04}$
	Rényi Divergence	$0.46_{\pm 0.02}$	$0.05_{\pm 0.03}$	$0.35_{\pm 0.03}$	$0.56_{\pm 0.02}$	$0.05_{\pm 0.03}$	$0.22_{\pm 0.04}$	$0.52_{\pm 0.02}$	$0.05_{\pm 0.02}$	$0.25_{\pm 0.05}$	$0.59_{\pm 0.03}$	$0.12_{\pm 0.04}$	$0.21_{\pm 0.04}$
	Fisher–Rao Distance	$0.50_{\pm 0.02}$	$0.12_{\pm 0.03}$	$0.30_{\pm 0.03}$	$0.56_{\pm 0.02}$	$0.06_{\pm 0.03}$	$0.19_{\pm 0.04}$	$0.52_{\pm 0.02}$	$0.03_{\pm 0.02}$	$0.24_{\pm 0.06}$	$0.61_{\pm 0.03}$	$0.17_{\pm 0.04}$	$0.21_{\pm 0.04}$
	Token SAR	$0.56_{\pm 0.02}$	$0.21_{\pm 0.03}$	$0.23_{\pm 0.03}$	$0.75_{\pm 0.02}$	$0.45_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.53_{\pm 0.02}$	$0.04_{\pm 0.02}$	$0.21_{\pm 0.05}$	$0.55_{\pm 0.03}$	$0.06_{\pm 0.03}$	$0.24_{\pm 0.06}$
	Claim-Conditioned Probability (CCP)	$0.76_{\pm 0.01}$	$\mathbf{0.56_{\pm 0.02}}$	$0.06_{\pm 0.02}$	$0.75_{\pm 0.02}$	$0.46_{\pm 0.03}$	$0.10_{\pm 0.02}$	$0.59_{\pm 0.02}$	$0.10_{\pm 0.02}$	$0.16_{\pm 0.03}$	$0.71_{\pm 0.03}$	$0.33_{\pm 0.04}$	$\mathbf{0.16_{\pm 0.04}}$
Sampling-based	Monte-Carlo Seq. Entropy	$0.34_{\pm 0.01}$	$-0.21_{\pm 0.03}$	$0.47_{\pm 0.02}$	$0.73_{\pm 0.02}$	$0.43_{\pm 0.03}$	$0.09_{\pm 0.03}$	$0.52_{\pm 0.02}$	$0.01_{\pm 0.02}$	$0.32_{\pm 0.06}$	$0.66_{\pm 0.03}$	$0.25_{\pm 0.04}$	$0.16_{\pm 0.04}$
	Monte-Carlo Norm. Seq. Entropy	$0.34_{\pm 0.01}$	$-0.20_{\pm 0.03}$	$0.48_{\pm 0.01}$	$0.73_{\pm 0.02}$	$0.39_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.51_{\pm 0.02}$	$0.02_{\pm 0.02}$	$0.31_{\pm 0.05}$	$0.60_{\pm 0.03}$	$0.14_{\pm 0.04}$	$0.20_{\pm 0.05}$
	Semantic Entropy	$0.32_{\pm 0.01}$	$-0.24_{\pm 0.03}$	$0.49_{\pm 0.01}$	$0.74_{\pm 0.02}$	$0.45_{\pm 0.03}$	$0.12_{\pm 0.03}$	$0.53_{\pm 0.02}$	$0.03_{\pm 0.02}$	$0.28_{\pm 0.06}$	$0.66_{\pm 0.03}$	$0.24_{\pm 0.04}$	$0.16_{\pm 0.05}$
	Semantic Density	$0.39_{\pm 0.01}$	$-0.13_{\pm 0.03}$	$0.45_{\pm 0.02}$	$0.79_{\pm 0.02}$	$0.53_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.57_{\pm 0.02}$	$0.10_{\pm 0.02}$	$0.15_{\pm 0.04}$	$0.43_{\pm 0.03}$	$-0.06_{\pm 0.03}$	$0.41_{\pm 0.04}$
	Sentence SAR	$0.40_{\pm 0.02}$	$-0.08_{\pm 0.03}$	$0.43_{\pm 0.02}$	$0.77_{\pm 0.02}$	$0.53_{\pm 0.03}$	$0.10_{\pm 0.03}$	$0.50_{\pm 0.02}$	$-0.02_{\pm 0.02}$	$0.35_{\pm 0.05}$	$\mathbf{0.72_{\pm 0.03}}$	$0.35_{\pm 0.04}$	$0.16_{\pm 0.04}$
	SAR	$0.35_{\pm 0.01}$	$-0.20_{\pm 0.03}$	$0.49_{\pm 0.02}$	$0.76_{\pm 0.02}$	$0.49_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.56_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.17_{\pm 0.04}$	$0.53_{\pm 0.02}$	$0.02_{\pm 0.03}$	$0.28_{\pm 0.06}$
	Cocoa MSP	$0.75_{\pm 0.01}$	$0.50_{\pm 0.02}$	$\mathbf{0.05_{\pm 0.02}}$	$0.78_{\pm 0.02}$	$0.52_{\pm 0.03}$	$0.11_{\pm 0.02}$	$0.56_{\pm 0.02}$	$0.07_{\pm 0.02}$	$0.21_{\pm 0.04}$	$0.62_{\pm 0.03}$	$0.12_{\pm 0.03}$	$0.16_{\pm 0.04}$
	Cocoa PPL	$0.55_{\pm 0.02}$	$0.21_{\pm 0.03}$	$0.24_{\pm 0.03}$	$0.75_{\pm 0.02}$	$0.44_{\pm 0.03}$	$0.09_{\pm 0.02}$	$0.57_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.17_{\pm 0.04}$	$0.50_{\pm 0.03}$	$-0.02_{\pm 0.03}$	$0.33_{\pm 0.06}$
	Cocoa MTE	$0.54_{\pm 0.02}$	$0.17_{\pm 0.03}$	$0.26_{\pm 0.03}$	$0.75_{\pm 0.02}$	$0.46_{\pm 0.03}$	$0.09_{\pm 0.03}$	$0.57_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.17_{\pm 0.04}$	$0.51_{\pm 0.03}$	$-0.02_{\pm 0.03}$	$0.31_{\pm 0.06}$
Internal	Attention Score	$0.73_{\pm 0.01}$	$0.47_{\pm 0.03}$	$0.12_{\pm 0.02}$	$0.63_{\pm 0.02}$	$0.18_{\pm 0.03}$	$0.14_{\pm 0.03}$	$0.48_{\pm 0.02}$	$-0.07_{\pm 0.02}$	$0.41_{\pm 0.05}$	$0.70_{\pm 0.03}$	$\mathbf{0.39_{\pm 0.05}}$	$0.20_{\pm 0.05}$
	RAUQ	$0.56_{\pm 0.02}$	$0.22_{\pm 0.03}$	$0.22_{\pm 0.03}$	$0.76_{\pm 0.02}$	$0.48_{\pm 0.03}$	$0.08_{\pm 0.03}$	$0.52_{\pm 0.02}$	$0.02_{\pm 0.02}$	$0.23_{\pm 0.05}$	$0.56_{\pm 0.03}$	$0.05_{\pm 0.03}$	$0.23_{\pm 0.05}$
	CSL	$0.42_{\pm 0.02}$	$-0.02_{\pm 0.03}$	$0.41_{\pm 0.02}$	$0.69_{\pm 0.02}$	$0.34_{\pm 0.03}$	$0.11_{\pm 0.03}$	$0.49_{\pm 0.02}$	$0.01_{\pm 0.02}$	$0.29_{\pm 0.05}$	$0.48_{\pm 0.03}$	$-0.02_{\pm 0.03}$	$0.38_{\pm 0.05}$
	EigenScore	$0.32_{\pm 0.01}$	$-0.25_{\pm 0.03}$	$0.49_{\pm 0.01}$	$0.68_{\pm 0.02}$	$0.39_{\pm 0.03}$	$0.19_{\pm 0.05}$	$0.48_{\pm 0.02}$	$-0.03_{\pm 0.02}$	$0.41_{\pm 0.04}$	$0.66_{\pm 0.03}$	$0.29_{\pm 0.05}$	$0.19_{\pm 0.05}$
Training-based	Mahalanobis Distance (MD)	$0.24_{\pm 0.01}$	$-0.45_{\pm 0.02}$	$0.49_{\pm 0.01}$	$0.52_{\pm 0.02}$	$0.02_{\pm 0.03}$	$0.28_{\pm 0.04}$	$0.45_{\pm 0.02}$	$-0.03_{\pm 0.02}$	$0.39_{\pm 0.05}$	$0.26_{\pm 0.02}$	$-0.21_{\pm 0.02}$	$0.49_{\pm 0.02}$
	Relative MD (RMD)	$0.29_{\pm 0.01}$	$-0.37_{\pm 0.02}$	$0.49_{\pm 0.01}$	$0.49_{\pm 0.02}$	$-0.05_{\pm 0.02}$	$0.32_{\pm 0.04}$	$0.49_{\pm 0.02}$	$0.01_{\pm 0.02}$	$0.32_{\pm 0.05}$	$0.21_{\pm 0.02}$	$-0.24_{\pm 0.02}$	$0.50_{\pm 0.02}$
	Robust Density Estimation (RDE)	$0.39_{\pm 0.01}$	$-0.31_{\pm 0.02}$	$0.42_{\pm 0.01}$	$0.38_{\pm 0.02}$	$-0.13_{\pm 0.03}$	$0.47_{\pm 0.02}$	$0.58_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.16_{\pm 0.04}$	$0.32_{\pm 0.02}$	$-0.19_{\pm 0.02}$	$0.50_{\pm 0.01}$
	HUQ-MD	$0.56_{\pm 0.02}$	$0.21_{\pm 0.03}$	$0.23_{\pm 0.03}$	$0.71_{\pm 0.02}$	$0.32_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.53_{\pm 0.02}$	$0.04_{\pm 0.02}$	$0.22_{\pm 0.05}$	$0.55_{\pm 0.03}$	$0.05_{\pm 0.03}$	$0.23_{\pm 0.06}$
	HUQ-RMD	$0.55_{\pm 0.02}$	$0.20_{\pm 0.03}$	$0.23_{\pm 0.03}$	$0.71_{\pm 0.02}$	$0.32_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.53_{\pm 0.02}$	$0.04_{\pm 0.02}$	$0.22_{\pm 0.05}$	$0.55_{\pm 0.03}$	$0.05_{\pm 0.03}$	$0.23_{\pm 0.06}$
Reflexive	P(True)	$0.48_{\pm 0.01}$	$0.03_{\pm 0.03}$	$0.33_{\pm 0.03}$	$0.48_{\pm 0.02}$	$-0.06_{\pm 0.02}$	$0.36_{\pm 0.04}$	$0.48_{\pm 0.02}$	$-0.02_{\pm 0.02}$	$0.40_{\pm 0.05}$	$0.26_{\pm 0.03}$	$-0.18_{\pm 0.03}$	$0.39_{\pm 0.05}$
	P(True) Sampling	$0.37_{\pm 0.01}$	$-0.21_{\pm 0.03}$	$0.48_{\pm 0.02}$	$0.54_{\pm 0.02}$	$0.03_{\pm 0.03}$	$0.22_{\pm 0.05}$	–	–	–	–	–	–
	P(True) Empirical	$0.51_{\pm 0.00}$	$0.02_{\pm 0.03}$	$0.31_{\pm 0.07}$	$0.51_{\pm 0.01}$	$0.02_{\pm 0.03}$	$0.32_{\pm 0.07}$	$0.50_{\pm 0.00}$	$-0.01_{\pm 0.02}$	$0.34_{\pm 0.07}$	$0.49_{\pm 0.00}$	$-0.02_{\pm 0.03}$	$0.35_{\pm 0.07}$
Black-box	NumSet	$0.48_{\pm 0.01}$	$-0.02_{\pm 0.03}$	$0.37_{\pm 0.06}$	$0.78_{\pm 0.01}$	$0.46_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.53_{\pm 0.01}$	$0.04_{\pm 0.02}$	$0.25_{\pm 0.06}$	$0.44_{\pm 0.02}$	$-0.05_{\pm 0.03}$	$0.45_{\pm 0.05}$
	LabelProb	$0.31_{\pm 0.01}$	$-0.34_{\pm 0.02}$	$0.50_{\pm 0.01}$	$0.76_{\pm 0.02}$	$0.49_{\pm 0.03}$	$0.09_{\pm 0.03}$	$0.52_{\pm 0.02}$	$0.04_{\pm 0.02}$	$0.25_{\pm 0.06}$	$0.59_{\pm 0.02}$	$0.11_{\pm 0.04}$	$0.21_{\pm 0.05}$
	Kernel Language Entropy (KLE)	$0.29_{\pm 0.01}$	$-0.38_{\pm 0.02}$	$0.50_{\pm 0.01}$	$0.80_{\pm 0.02}$	$0.53_{\pm 0.03}$	$\mathbf{0.06_{\pm 0.02}}$	$0.58_{\pm 0.02}$	$0.10_{\pm 0.02}$	$0.15_{\pm 0.04}$	$0.50_{\pm 0.03}$	$0.02_{\pm 0.03}$	$0.34_{\pm 0.05}$
	EigValLap NLI (entail)	$0.29_{\pm 0.01}$	$-0.37_{\pm 0.02}$	$0.49_{\pm 0.01}$	$\mathbf{0.81_{\pm 0.01}}$	$\mathbf{0.57_{\pm 0.03}}$	$0.08_{\pm 0.02}$	$0.56_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.16_{\pm 0.04}$	$0.53_{\pm 0.03}$	$0.02_{\pm 0.03}$	$0.29_{\pm 0.05}$
	EigValLap NLI (contra)	$0.37_{\pm 0.01}$	$-0.28_{\pm 0.02}$	$0.47_{\pm 0.02}$	$0.77_{\pm 0.02}$	$0.49_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.61_{\pm 0.02}$	$\mathbf{0.19_{\pm 0.02}}$	$0.11_{\pm 0.03}$	$0.47_{\pm 0.03}$	$-0.04_{\pm 0.03}$	$0.39_{\pm 0.05}$
	EigValLap Jaccard	$0.30_{\pm 0.01}$	$-0.27_{\pm 0.03}$	$0.50_{\pm 0.01}$	$0.69_{\pm 0.02}$	$0.31_{\pm 0.03}$	$0.09_{\pm 0.03}$	$0.55_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.16_{\pm 0.04}$	$0.43_{\pm 0.03}$	$-0.07_{\pm 0.03}$	$0.42_{\pm 0.04}$
	Eccentricity NLI (entail)	$0.29_{\pm 0.01}$	$-0.38_{\pm 0.02}$	$0.49_{\pm 0.01}$	$0.77_{\pm 0.02}$	$0.51_{\pm 0.03}$	$0.12_{\pm 0.03}$	$0.55_{\pm 0.02}$	$0.07_{\pm 0.02}$	$0.19_{\pm 0.05}$	$0.57_{\pm 0.03}$	$0.10_{\pm 0.04}$	$0.25_{\pm 0.05}$
	Eccentricity NLI (contra)	$0.37_{\pm 0.01}$	$-0.28_{\pm 0.02}$	$0.48_{\pm 0.02}$	$0.80_{\pm 0.02}$	$0.53_{\pm 0.03}$	$0.07_{\pm 0.02}$	$\mathbf{0.62_{\pm 0.02}}$	$0.19_{\pm 0.02}$	$0.12_{\pm 0.03}$	$0.46_{\pm 0.03}$	$-0.04_{\pm 0.03}$	$0.42_{\pm 0.05}$
	Eccentricity Jaccard	$0.48_{\pm 0.01}$	$-0.02_{\pm 0.03}$	$0.38_{\pm 0.06}$	$0.72_{\pm 0.02}$	$0.39_{\pm 0.03}$	$0.11_{\pm 0.03}$	$0.50_{\pm 0.02}$	$-0.02_{\pm 0.02}$	$0.39_{\pm 0.06}$	$0.51_{\pm 0.02}$	$0.01_{\pm 0.03}$	$0.31_{\pm 0.07}$
	DegMat NLI (entail)	$0.29_{\pm 0.01}$	$-0.38_{\pm 0.02}$	$0.49_{\pm 0.01}$	$0.80_{\pm 0.01}$	$0.54_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.55_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.17_{\pm 0.05}$	$0.55_{\pm 0.03}$	$0.05_{\pm 0.03}$	$0.26_{\pm 0.05}$
	DegMat NLI (contra)	$0.34_{\pm 0.01}$	$-0.32_{\pm 0.02}$	$0.49_{\pm 0.01}$	$0.77_{\pm 0.02}$	$0.46_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.61_{\pm 0.02}$	$0.19_{\pm 0.02}$	$0.11_{\pm 0.03}$	$0.46_{\pm 0.03}$	$-0.04_{\pm 0.03}$	$0.40_{\pm 0.05}$
	DegMat Jaccard	$0.30_{\pm 0.01}$	$-0.26_{\pm 0.03}$	$0.50_{\pm 0.01}$	$0.70_{\pm 0.02}$	$0.36_{\pm 0.03}$	$0.09_{\pm 0.03}$	$0.55_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.16_{\pm 0.04}$	$0.44_{\pm 0.03}$	$-0.06_{\pm 0.03}$	$0.41_{\pm 0.04}$
	LUQ	$0.31_{\pm 0.01}$	$-0.36_{\pm 0.02}$	$0.49_{\pm 0.01}$	$0.80_{\pm 0.02}$	$0.53_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.61_{\pm 0.02}$	$0.18_{\pm 0.02}$	$\mathbf{0.10_{\pm 0.03}}$	$0.46_{\pm 0.03}$	$-0.04_{\pm 0.03}$	$0.38_{\pm 0.05}$
	Lexical Sim. (ROUGE-L)	$0.29_{\pm 0.01}$	$-0.28_{\pm 0.03}$	$0.50_{\pm 0.01}$	$0.70_{\pm 0.02}$	$0.36_{\pm 0.03}$	$0.09_{\pm 0.03}$	$0.53_{\pm 0.02}$	$0.06_{\pm 0.02}$	$0.18_{\pm 0.05}$	$0.39_{\pm 0.02}$	$-0.12_{\pm 0.02}$	$0.47_{\pm 0.03}$
	Lexical Sim. (BLEU)	$0.31_{\pm 0.01}$	$-0.25_{\pm 0.03}$	$0.49_{\pm 0.01}$	$0.70_{\pm 0.02}$	$0.36_{\pm 0.03}$	$0.10_{\pm 0.03}$	$0.54_{\pm 0.02}$	$0.07_{\pm 0.02}$	$0.17_{\pm 0.05}$	$0.34_{\pm 0.02}$	$-0.16_{\pm 0.02}$	$0.49_{\pm 0.02}$
AlignScore	AS-Precision(gold_answer)	$\mathbf{0.70_{\pm 0.01}}$	$\mathbf{0.51_{\pm 0.03}}$	$\mathbf{0.15_{\pm 0.01}}$	$0.73_{\pm 0.02}$	$0.60_{\pm 0.03}$	$0.30_{\pm 0.04}$	$\mathbf{0.61_{\pm 0.02}}$	$\mathbf{0.20_{\pm 0.02}}$	$\mathbf{0.10_{\pm 0.03}}$	$0.71_{\pm 0.03}$	$0.33_{\pm 0.05}$	$\mathbf{0.16_{\pm 0.04}}$
	AS-Precision(wikipage)	–	–	–	$0.86_{\pm 0.01}$	$0.66_{\pm 0.03}$	$\mathbf{0.07_{\pm 0.02}}$	$0.60_{\pm 0.02}$	$0.16_{\pm 0.02}$	$0.13_{\pm 0.03}$	–	–	–
	AS-Recall(gold_answer)	$0.61_{\pm 0.01}$	$0.27_{\pm 0.03}$	$0.16_{\pm 0.02}$	$\mathbf{0.91_{\pm 0.01}}$	$\mathbf{0.84_{\pm 0.02}}$	$0.19_{\pm 0.05}$	$0.59_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.16_{\pm 0.04}$	$\mathbf{0.72_{\pm 0.03}}$	$\mathbf{0.47_{\pm 0.05}}$	$0.18_{\pm 0.05}$
	AS-Recall(wikipage)	–	–	–	$0.50_{\pm 0.02}$	$0.01_{\pm 0.03}$	$0.35_{\pm 0.06}$	$0.56_{\pm 0.02}$	$0.07_{\pm 0.02}$	$0.19_{\pm 0.05}$	–	–	–

Table 8: AUROC ↑, PRR ↑, RCE ↓ for **Llama-2-7B-Chat** across tasks (higher AUROC/PRR and lower RCE are better; color shows numerical value). Uncertainty estimators and AlignScore variants are ranked separately: bold = best per column (ties marked); underline = second-best (ties marked) within the uncertainty group; AlignScore group shows only the best.

	Estimator	RAGTruth			PreciseWikiQA			LongWiki			NonExistentRefusal		
		AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓
Baseline	Output Token Count	0.60±0.01	0.25±0.03	0.16±0.03	0.61±0.02	0.10±0.03	0.17±0.04	0.48±0.01	−0.10±0.02	0.39±0.05	0.57±0.02	0.16±0.06	0.28±0.06
Information-based	Maximum Seq. Probability (MSP)	0.53±0.01	0.12±0.03	0.30±0.04	0.73±0.02	0.35±0.04	0.14±0.03	0.56±0.01	0.05±0.03	0.18±0.03	0.57±0.04	0.12±0.05	0.28±0.06
	Perplexity (PPL)	0.41±0.01	−0.10±0.03	0.47±0.03	0.65±0.02	0.17±0.04	0.12±0.03	0.59±0.01	0.14±0.03	0.12±0.03	0.49±0.04	0.02±0.04	0.34±0.05
	Mean Token Entropy (MTE)	0.53±0.01	0.13±0.03	0.27±0.04	0.60±0.02	0.10±0.03	0.19±0.04	0.59±0.01	0.15±0.03	0.12±0.03	0.51±0.04	0.02±0.04	0.33±0.06
	Pointwise Mutual Info. (PMI)	0.32±0.01	−0.32±0.03	0.50±0.01	0.45±0.03	−0.02±0.04	0.40±0.04	0.53±0.01	0.04±0.02	0.27±0.05	0.46±0.03	−0.07±0.02	0.35±0.04
	Conditional PMI	0.55±0.01	0.17±0.03	0.25±0.04	0.31±0.02	−0.19±0.02	0.49±0.01	0.41±0.01	−0.21±0.02	0.49±0.01	0.51±0.04	0.08±0.05	0.32±0.05
	Self Certainty	0.46±0.01	0.04±0.03	0.39±0.02	0.56±0.02	0.03±0.03	0.27±0.04	0.57±0.01	0.11±0.02	0.13±0.03	0.52±0.04	0.03±0.04	0.30±0.05
	Rényi Divergence	0.46±0.01	0.02±0.03	0.42±0.03	0.54±0.02	0.02±0.03	0.29±0.04	0.56±0.01	0.09±0.02	0.14±0.04	0.51±0.04	0.02±0.04	0.28±0.05
	Fisher–Rao Distance	0.46±0.01	0.06±0.03	0.39±0.02	0.54±0.02	0.02±0.03	0.28±0.04	0.57±0.01	0.12±0.02	0.12±0.03	0.54±0.04	0.07±0.04	0.27±0.04
	Token SAR	0.41±0.01	−0.10±0.03	0.47±0.03	0.72±0.02	0.30±0.04	0.11±0.03	0.59±0.01	0.14±0.03	0.12±0.03	0.50±0.04	0.03±0.04	0.32±0.06
	Claim-Conditioned Probability (CCP)	0.49±0.01	0.00±0.03	0.36±0.03	0.72±0.02	0.30±0.04	0.12±0.03	0.63±0.01	0.19±0.02	0.09±0.02	0.64±0.04	**0.19±0.05**	0.20±0.05
Sampling-based	Monte-Carlo Seq. Entropy	0.46±0.01	0.03±0.03	0.37±0.03	0.70±0.02	0.27±0.04	0.11±0.03	0.51±0.01	0.00±0.03	0.27±0.05	0.56±0.04	0.11±0.05	0.30±0.05
	Monte-Carlo Norm. Seq. Entropy	0.38±0.01	−0.17±0.03	0.49±0.01	0.68±0.02	0.22±0.04	0.12±0.03	0.54±0.01	0.09±0.03	0.18±0.05	0.44±0.04	−0.01±0.04	0.42±0.05
	Semantic Entropy	0.46±0.01	0.03±0.03	0.38±0.02	0.71±0.02	0.31±0.04	0.12±0.03	0.51±0.01	0.02±0.03	0.26±0.05	0.55±0.04	0.09±0.04	0.29±0.05
	Semantic Density	0.48±0.01	0.02±0.03	0.38±0.05	0.72±0.03	0.36±0.04	0.18±0.05	0.58±0.01	0.16±0.02	0.11±0.03	0.49±0.03	0.00±0.04	0.36±0.06
	Sentence SAR	0.56±0.01	0.19±0.03	0.26±0.02	0.73±0.02	0.37±0.04	0.15±0.04	0.49±0.01	−0.03±0.02	0.38±0.05	**0.64±0.03**	0.13±0.04	**0.16±0.04**
	SAR	0.42±0.01	−0.07±0.03	0.44±0.02	0.73±0.02	0.35±0.04	0.13±0.03	0.61±0.02	0.27±0.02	0.13±0.02	0.39±0.03	−0.10±0.03	0.45±0.04
	Cocoa MSP	0.50±0.01	0.07±0.03	0.35±0.04	**0.75±0.02**	**0.39±0.04**	0.13±0.04	0.63±0.01	0.22±0.03	0.08±0.02	0.49±0.04	0.04±0.05	0.36±0.05
	Cocoa PPL	0.42±0.01	−0.09±0.03	0.46±0.02	0.70±0.02	0.26±0.04	0.12±0.03	0.65±0.01	0.30±0.02	**0.06±0.02**	0.44±0.04	−0.03±0.04	0.41±0.05
	Cocoa MTE	0.50±0.01	0.07±0.03	0.33±0.04	0.68±0.02	0.21±0.04	**0.11±0.03**	0.65±0.01	0.30±0.02	0.07±0.02	0.45±0.04	−0.03±0.04	0.39±0.05
Internal	Attention Score	**0.65±0.01**	**0.31±0.03**	**0.14±0.02**	0.58±0.02	0.06±0.03	0.22±0.05	0.45±0.01	−0.14±0.02	0.47±0.03	0.60±0.04	0.17±0.05	0.23±0.06
	RAUQ	0.42±0.01	−0.09±0.03	0.46±0.03	0.71±0.02	0.33±0.04	0.13±0.03	0.58±0.01	0.13±0.03	0.13±0.03	0.50±0.04	0.02±0.04	0.33±0.05
	CSL	0.37±0.01	−0.22±0.03	0.50±0.01	0.58±0.03	0.09±0.04	0.19±0.05	0.56±0.02	0.08±0.03	0.18±0.04	0.45±0.03	−0.05±0.03	0.38±0.05
	EigenScore	0.48±0.01	0.04±0.03	0.37±0.03	0.63±0.03	0.22±0.04	0.23±0.05	0.42±0.01	−0.15±0.02	0.49±0.01	0.59±0.04	0.16±0.05	0.28±0.04
Training-based	Mahalanobis Distance (MD)	0.33±0.01	−0.38±0.02	0.48±0.01	0.50±0.03	0.00±0.03	0.33±0.06	0.40±0.01	−0.16±0.02	0.49±0.02	0.18±0.03	−0.22±0.02	0.50±0.03
	Relative MD (RMD)	0.33±0.01	−0.37±0.02	0.48±0.01	0.50±0.03	0.01±0.03	0.33±0.06	0.42±0.01	−0.14±0.02	0.48±0.02	0.18±0.03	−0.22±0.02	0.51±0.03
	Robust Density Estimation (RDE)	0.33±0.01	−0.36±0.02	0.46±0.02	0.57±0.02	0.07±0.03	0.23±0.05	0.44±0.01	−0.09±0.02	0.48±0.02	0.25±0.02	−0.20±0.02	0.51±0.02
	HUQ-MD	0.41±0.01	−0.10±0.03	0.47±0.03	0.65±0.02	0.17±0.04	0.11±0.03	0.59±0.01	0.14±0.03	0.12±0.03	0.49±0.04	0.02±0.04	0.34±0.05
	HUQ-RMD	0.41±0.01	−0.10±0.03	0.47±0.03	0.65±0.02	0.17±0.04	0.11±0.03	0.59±0.01	0.14±0.03	0.12±0.03	0.49±0.04	0.02±0.04	0.34±0.05
Reflexive	P(True)	0.43±0.01	−0.15±0.03	0.46±0.02	0.56±0.02	0.04±0.03	0.24±0.05	0.49±0.02	−0.04±0.02	0.36±0.04	0.39±0.04	−0.07±0.04	0.43±0.04
	P(True) Sampling	0.50±0.01	0.04±0.03	0.33±0.05	0.53±0.03	0.04±0.03	0.28±0.06	–	–	–	–	–	–
	P(True) Empirical	0.52±0.01	0.06±0.03	0.24±0.05	0.48±0.02	−0.03±0.03	0.36±0.05	0.49±0.02	0.00±0.03	0.39±0.05	0.54±0.03	0.02±0.03	0.28±0.07
Black-box	NumSet	0.48±0.01	−0.03±0.03	0.37±0.06	0.72±0.02	0.32±0.04	0.11±0.04	0.51±0.01	0.02±0.02	0.28±0.06	0.45±0.03	−0.03±0.03	0.39±0.06
	LabelProb	0.52±0.01	0.05±0.03	0.27±0.05	0.70±0.02	0.33±0.04	0.14±0.04	0.52±0.01	0.09±0.02	0.27±0.05	0.49±0.03	−0.02±0.03	0.35±0.06
	Kernel Language Entropy (KLE)	0.47±0.01	0.00±0.03	0.39±0.03	0.70±0.03	0.35±0.04	0.18±0.04	0.58±0.02	0.18±0.02	0.13±0.03	0.50±0.04	0.00±0.03	0.34±0.05
	EigValLap NLI (entail)	0.47±0.01	−0.02±0.03	0.38±0.04	0.71±0.03	0.36±0.04	0.16±0.04	0.57±0.02	0.16±0.02	0.15±0.04	0.47±0.04	−0.03±0.03	0.37±0.06
	EigValLap NLI (contra)	0.43±0.01	−0.17±0.03	0.43±0.03	0.61±0.03	0.20±0.04	0.24±0.05	**0.65±0.01**	0.31±0.02	0.07±0.02	0.53±0.04	0.02±0.03	0.30±0.06
	EigValLap Jaccard	0.37±0.01	−0.17±0.03	0.49±0.01	0.64±0.03	0.20±0.04	0.20±0.04	0.60±0.01	0.25±0.02	0.08±0.02	0.32±0.03	−0.14±0.03	0.49±0.02
	Eccentricity NLI (entail)	0.51±0.01	0.04±0.03	0.31±0.05	0.72±0.03	0.36±0.04	0.16±0.04	0.57±0.02	0.17±0.02	0.18±0.04	0.49±0.03	−0.03±0.03	0.34±0.05
	Eccentricity NLI (contra)	0.43±0.01	−0.20±0.02	0.44±0.03	0.70±0.02	0.32±0.04	0.13±0.04	0.64±0.01	0.28±0.02	0.06±0.02	0.54±0.04	0.02±0.03	0.28±0.05
	Eccentricity Jaccard	0.56±0.01	0.15±0.03	0.22±0.05	0.66±0.03	0.26±0.04	0.20±0.04	0.48±0.01	−0.02±0.02	0.39±0.06	0.52±0.03	0.01±0.03	0.31±0.07
	DegMat NLI (entail)	0.48±0.01	0.01±0.03	0.35±0.04	0.72±0.03	0.37±0.04	0.18±0.05	0.55±0.02	0.13±0.02	0.20±0.05	0.48±0.04	−0.03±0.03	0.37±0.06
	DegMat NLI (contra)	0.41±0.01	−0.23±0.02	0.46±0.03	0.63±0.03	0.22±0.04	0.21±0.05	0.65±0.01	0.29±0.02	0.07±0.02	0.53±0.04	0.02±0.03	0.30±0.07
	DegMat Jaccard	0.40±0.01	−0.08±0.03	0.45±0.02	0.66±0.03	0.23±0.04	0.18±0.04	0.61±0.01	0.25±0.02	0.08±0.02	0.33±0.03	−0.14±0.03	0.49±0.02
	LUQ	0.37±0.01	−0.24±0.03	0.49±0.02	0.70±0.03	0.35±0.04	0.18±0.05	0.65±0.01	**0.31±0.02**	0.08±0.02	0.49±0.04	−0.01±0.03	0.35±0.05
	Lexical Sim. (ROUGE-L)	0.40±0.01	−0.08±0.03	0.44±0.02	0.65±0.03	0.22±0.04	0.18±0.04	0.56±0.01	0.14±0.02	0.11±0.03	0.25±0.03	−0.19±0.02	0.51±0.02
	Lexical Sim. (BLEU)	0.42±0.01	−0.05±0.03	0.43±0.02	0.65±0.03	0.22±0.04	0.15±0.04	0.60±0.01	0.23±0.02	0.07±0.02	0.25±0.03	−0.18±0.03	0.51±0.02
AlignScore	AS-Precision(gold_answer)	**0.71±0.01**	**0.47±0.03**	**0.06±0.02**	0.88±0.02	0.74±0.04	0.21±0.05	0.65±0.01	0.33±0.02	0.06±0.02	0.55±0.04	0.09±0.05	**0.29±0.05**
	AS-Precision(wikipage)	–	–	–	0.77±0.02	0.32±0.04	**0.10±0.03**	**0.70±0.01**	**0.41±0.02**	**0.05±0.01**	–	–	–
	AS-Recall(gold_answer)	0.57±0.01	0.17±0.03	0.20±0.04	**0.91±0.02**	**0.80±0.03**	0.29±0.05	0.58±0.01	0.19±0.03	0.12±0.03	**0.61±0.04**	**0.26±0.07**	0.30±0.05
	AS-Recall(wikipage)	–	–	–	0.53±0.03	0.04±0.03	0.28±0.06	0.57±0.01	0.15±0.02	0.13±0.04	–	–	–

Table 9: AUROC ↑, PRR ↑, RCE ↓ for **Llama-2-13B-Chat** across tasks (higher AUROC/PRR and lower RCE are better; color shows numerical value). Uncertainty estimators and AlignScore variants are ranked separately: bold = best per column (ties marked); underline = second-best (ties marked) within the uncertainty group; AlignScore group shows only the best.

	Estimator	RAGTruth			PreciseWikiQA			LongWiki			NonExistentRefusal		
		AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓	AUROC ↑	PRR ↑	RCE ↓
Baseline	Output Token Count	**0.75**$_{\pm 0.01}$	**0.49**$_{\pm 0.03}$	**0.05**$_{\pm 0.02}$	0.59$_{\pm 0.02}$	0.10$_{\pm 0.03}$	0.18$_{\pm 0.04}$	0.50$_{\pm 0.00}$	0.00$_{\pm 0.03}$	0.34$_{\pm 0.07}$	0.87$_{\pm 0.01}$	0.74$_{\pm 0.03}$	0.22$_{\pm 0.05}$
Information-based	Maximum Seq. Probability (MSP)	0.63$_{\pm 0.01}$	0.32$_{\pm 0.03}$	0.12$_{\pm 0.02}$	0.64$_{\pm 0.02}$	0.24$_{\pm 0.03}$	0.16$_{\pm 0.03}$	0.59$_{\pm 0.02}$	0.13$_{\pm 0.03}$	0.12$_{\pm 0.03}$	0.76$_{\pm 0.02}$	0.55$_{\pm 0.03}$	0.09$_{\pm 0.03}$
	Perplexity (PPL)	0.35$_{\pm 0.01}$	−0.22$_{\pm 0.03}$	0.49$_{\pm 0.01}$	0.55$_{\pm 0.02}$	0.08$_{\pm 0.03}$	0.27$_{\pm 0.04}$	0.59$_{\pm 0.02}$	0.13$_{\pm 0.03}$	0.12$_{\pm 0.03}$	0.25$_{\pm 0.02}$	−0.25$_{\pm 0.03}$	0.50$_{\pm 0.01}$
	Mean Token Entropy (MTE)	0.51$_{\pm 0.01}$	0.18$_{\pm 0.03}$	0.34$_{\pm 0.02}$	0.55$_{\pm 0.02}$	0.08$_{\pm 0.03}$	0.27$_{\pm 0.04}$	0.59$_{\pm 0.02}$	0.12$_{\pm 0.03}$	0.13$_{\pm 0.04}$	0.26$_{\pm 0.02}$	−0.24$_{\pm 0.03}$	0.50$_{\pm 0.01}$
	Pointwise Mutual Info. (PMI)	0.31$_{\pm 0.01}$	−0.35$_{\pm 0.03}$	0.49$_{\pm 0.02}$	0.47$_{\pm 0.02}$	0.00$_{\pm 0.03}$	0.39$_{\pm 0.05}$	0.49$_{\pm 0.02}$	0.00$_{\pm 0.02}$	0.34$_{\pm 0.04}$	0.42$_{\pm 0.02}$	−0.17$_{\pm 0.02}$	0.39$_{\pm 0.02}$
	Conditional PMI	0.56$_{\pm 0.01}$	0.23$_{\pm 0.03}$	0.23$_{\pm 0.03}$	0.35$_{\pm 0.02}$	−0.20$_{\pm 0.02}$	0.44$_{\pm 0.03}$	0.40$_{\pm 0.02}$	−0.18$_{\pm 0.02}$	0.49$_{\pm 0.02}$	0.75$_{\pm 0.02}$	0.47$_{\pm 0.03}$	0.10$_{\pm 0.02}$
	Self Certainty	0.43$_{\pm 0.01}$	0.02$_{\pm 0.03}$	0.41$_{\pm 0.02}$	0.54$_{\pm 0.02}$	0.00$_{\pm 0.02}$	0.26$_{\pm 0.04}$	0.55$_{\pm 0.02}$	0.06$_{\pm 0.02}$	0.20$_{\pm 0.05}$	0.40$_{\pm 0.02}$	−0.08$_{\pm 0.03}$	0.35$_{\pm 0.05}$
	Rényi Divergence	0.42$_{\pm 0.01}$	−0.01$_{\pm 0.03}$	0.43$_{\pm 0.02}$	0.53$_{\pm 0.02}$	−0.01$_{\pm 0.02}$	0.28$_{\pm 0.05}$	0.54$_{\pm 0.02}$	0.04$_{\pm 0.02}$	0.23$_{\pm 0.05}$	0.41$_{\pm 0.02}$	−0.07$_{\pm 0.03}$	0.33$_{\pm 0.04}$
	Fisher–Rao Distance	0.44$_{\pm 0.01}$	0.03$_{\pm 0.03}$	0.41$_{\pm 0.02}$	0.50$_{\pm 0.02}$	−0.04$_{\pm 0.02}$	0.30$_{\pm 0.05}$	0.56$_{\pm 0.02}$	0.07$_{\pm 0.03}$	0.18$_{\pm 0.04}$	0.45$_{\pm 0.02}$	−0.01$_{\pm 0.03}$	0.27$_{\pm 0.03}$
	Token SAR	0.35$_{\pm 0.01}$	−0.22$_{\pm 0.03}$	0.50$_{\pm 0.01}$	0.63$_{\pm 0.02}$	0.22$_{\pm 0.03}$	0.17$_{\pm 0.04}$	0.60$_{\pm 0.02}$	0.15$_{\pm 0.02}$	0.11$_{\pm 0.03}$	0.21$_{\pm 0.02}$	−0.28$_{\pm 0.03}$	0.49$_{\pm 0.02}$
	Claim-Conditioned Probability (CCP)	0.58$_{\pm 0.01}$	0.23$_{\pm 0.03}$	0.15$_{\pm 0.04}$	0.70$_{\pm 0.02}$	0.30$_{\pm 0.03}$	0.09$_{\pm 0.02}$	0.64$_{\pm 0.02}$	0.28$_{\pm 0.02}$	0.06$_{\pm 0.02}$	0.85$_{\pm 0.01}$	0.71$_{\pm 0.02}$	**0.08**$_{\pm 0.03}$
Sampling-based	Monte-Carlo Seq. Entropy	0.46$_{\pm 0.01}$	0.06$_{\pm 0.03}$	0.36$_{\pm 0.02}$	0.57$_{\pm 0.02}$	0.16$_{\pm 0.04}$	0.21$_{\pm 0.05}$	0.54$_{\pm 0.02}$	0.07$_{\pm 0.02}$	0.22$_{\pm 0.05}$	0.52$_{\pm 0.02}$	0.02$_{\pm 0.03}$	0.26$_{\pm 0.06}$
	Monte-Carlo Norm. Seq. Entropy	0.38$_{\pm 0.01}$	−0.14$_{\pm 0.03}$	0.46$_{\pm 0.02}$	0.57$_{\pm 0.02}$	0.17$_{\pm 0.04}$	0.22$_{\pm 0.05}$	0.55$_{\pm 0.02}$	0.08$_{\pm 0.02}$	0.20$_{\pm 0.05}$	0.40$_{\pm 0.02}$	−0.15$_{\pm 0.02}$	0.47$_{\pm 0.02}$
	Semantic Entropy	0.48$_{\pm 0.01}$	0.08$_{\pm 0.03}$	0.36$_{\pm 0.03}$	0.59$_{\pm 0.02}$	0.18$_{\pm 0.04}$	0.19$_{\pm 0.05}$	0.58$_{\pm 0.02}$	0.14$_{\pm 0.02}$	0.15$_{\pm 0.04}$	0.51$_{\pm 0.02}$	0.01$_{\pm 0.03}$	0.31$_{\pm 0.06}$
	Semantic Density	0.49$_{\pm 0.01}$	0.02$_{\pm 0.03}$	0.36$_{\pm 0.05}$	**0.75**$_{\pm 0.02}$	**0.46**$_{\pm 0.03}$	**0.07**$_{\pm 0.02}$	0.62$_{\pm 0.02}$	0.24$_{\pm 0.02}$	0.09$_{\pm 0.02}$	0.33$_{\pm 0.02}$	−0.19$_{\pm 0.03}$	0.49$_{\pm 0.01}$
	Sentence SAR	0.58$_{\pm 0.01}$	0.26$_{\pm 0.03}$	0.24$_{\pm 0.02}$	0.63$_{\pm 0.02}$	0.24$_{\pm 0.03}$	0.16$_{\pm 0.03}$	0.54$_{\pm 0.02}$	0.03$_{\pm 0.02}$	0.22$_{\pm 0.05}$	0.61$_{\pm 0.02}$	0.15$_{\pm 0.03}$	0.13$_{\pm 0.03}$
	SAR	0.46$_{\pm 0.01}$	0.01$_{\pm 0.03}$	0.37$_{\pm 0.03}$	0.64$_{\pm 0.02}$	0.26$_{\pm 0.04}$	0.14$_{\pm 0.03}$	0.61$_{\pm 0.02}$	0.25$_{\pm 0.02}$	0.09$_{\pm 0.03}$	0.36$_{\pm 0.02}$	−0.21$_{\pm 0.02}$	0.49$_{\pm 0.02}$
	Cocoa MSP	0.60$_{\pm 0.01}$	0.28$_{\pm 0.03}$	0.16$_{\pm 0.04}$	0.68$_{\pm 0.02}$	0.30$_{\pm 0.03}$	0.11$_{\pm 0.03}$	0.64$_{\pm 0.02}$	0.26$_{\pm 0.02}$	0.09$_{\pm 0.02}$	0.68$_{\pm 0.02}$	0.37$_{\pm 0.03}$	0.13$_{\pm 0.04}$
	Cocoa PPL	0.38$_{\pm 0.01}$	−0.13$_{\pm 0.03}$	0.48$_{\pm 0.01}$	0.60$_{\pm 0.02}$	0.18$_{\pm 0.03}$	0.20$_{\pm 0.04}$	0.65$_{\pm 0.02}$	0.26$_{\pm 0.02}$	0.09$_{\pm 0.02}$	0.23$_{\pm 0.02}$	−0.28$_{\pm 0.03}$	0.50$_{\pm 0.01}$
	Cocoa MTE	0.48$_{\pm 0.01}$	0.13$_{\pm 0.03}$	0.36$_{\pm 0.02}$	0.61$_{\pm 0.02}$	0.18$_{\pm 0.03}$	0.17$_{\pm 0.04}$	0.64$_{\pm 0.02}$	0.26$_{\pm 0.02}$	0.07$_{\pm 0.02}$	0.24$_{\pm 0.02}$	−0.27$_{\pm 0.03}$	0.50$_{\pm 0.01}$
Internal	Attention Score	0.69$_{\pm 0.01}$	0.39$_{\pm 0.03}$	0.14$_{\pm 0.02}$	0.57$_{\pm 0.02}$	0.09$_{\pm 0.03}$	0.19$_{\pm 0.05}$	0.47$_{\pm 0.02}$	−0.07$_{\pm 0.03}$	0.47$_{\pm 0.03}$	**0.87**$_{\pm 0.01}$	**0.75**$_{\pm 0.02}$	0.17$_{\pm 0.03}$
	RAUQ	0.35$_{\pm 0.01}$	−0.20$_{\pm 0.03}$	0.49$_{\pm 0.01}$	0.56$_{\pm 0.02}$	0.17$_{\pm 0.04}$	0.31$_{\pm 0.04}$	0.58$_{\pm 0.02}$	0.11$_{\pm 0.02}$	0.13$_{\pm 0.04}$	0.27$_{\pm 0.02}$	−0.22$_{\pm 0.03}$	0.50$_{\pm 0.01}$
	CSL	0.38$_{\pm 0.01}$	−0.15$_{\pm 0.03}$	0.47$_{\pm 0.02}$	0.59$_{\pm 0.02}$	0.18$_{\pm 0.03}$	0.22$_{\pm 0.04}$	0.54$_{\pm 0.02}$	0.08$_{\pm 0.02}$	0.22$_{\pm 0.05}$	0.49$_{\pm 0.02}$	0.18$_{\pm 0.04}$	0.36$_{\pm 0.03}$
	EigenScore	0.51$_{\pm 0.01}$	0.12$_{\pm 0.03}$	0.35$_{\pm 0.02}$	0.53$_{\pm 0.02}$	0.09$_{\pm 0.03}$	0.28$_{\pm 0.05}$	0.50$_{\pm 0.02}$	−0.04$_{\pm 0.02}$	0.33$_{\pm 0.05}$	0.42$_{\pm 0.02}$	−0.15$_{\pm 0.02}$	0.46$_{\pm 0.03}$
Training-based	Mahalanobis Distance (MD)	0.23$_{\pm 0.01}$	−0.60$_{\pm 0.02}$	0.49$_{\pm 0.01}$	0.47$_{\pm 0.02}$	−0.09$_{\pm 0.02}$	0.36$_{\pm 0.05}$	0.35$_{\pm 0.02}$	−0.21$_{\pm 0.02}$	0.49$_{\pm 0.01}$	0.60$_{\pm 0.02}$	0.27$_{\pm 0.03}$	0.21$_{\pm 0.03}$
	Relative MD (RMD)	0.25$_{\pm 0.01}$	−0.48$_{\pm 0.03}$	0.49$_{\pm 0.01}$	0.45$_{\pm 0.02}$	−0.10$_{\pm 0.02}$	0.39$_{\pm 0.05}$	0.36$_{\pm 0.02}$	−0.17$_{\pm 0.02}$	0.49$_{\pm 0.01}$	0.52$_{\pm 0.02}$	0.07$_{\pm 0.03}$	0.29$_{\pm 0.05}$
	Robust Density Estimation (RDE)	0.25$_{\pm 0.01}$	−0.56$_{\pm 0.02}$	0.48$_{\pm 0.01}$	0.59$_{\pm 0.02}$	0.10$_{\pm 0.03}$	0.16$_{\pm 0.04}$	0.46$_{\pm 0.02}$	−0.11$_{\pm 0.02}$	0.44$_{\pm 0.04}$	0.52$_{\pm 0.02}$	0.08$_{\pm 0.03}$	0.31$_{\pm 0.02}$
	HUQ-MD	0.35$_{\pm 0.01}$	−0.22$_{\pm 0.03}$	0.49$_{\pm 0.01}$	0.54$_{\pm 0.02}$	0.08$_{\pm 0.03}$	0.28$_{\pm 0.04}$	0.59$_{\pm 0.02}$	0.13$_{\pm 0.03}$	0.12$_{\pm 0.03}$	0.25$_{\pm 0.02}$	−0.25$_{\pm 0.03}$	0.50$_{\pm 0.01}$
	HUQ-RMD	0.35$_{\pm 0.01}$	−0.21$_{\pm 0.03}$	0.49$_{\pm 0.01}$	0.54$_{\pm 0.02}$	0.08$_{\pm 0.03}$	0.28$_{\pm 0.04}$	0.59$_{\pm 0.02}$	0.13$_{\pm 0.03}$	0.12$_{\pm 0.03}$	0.25$_{\pm 0.02}$	−0.25$_{\pm 0.03}$	0.50$_{\pm 0.01}$
Reflexive	P(True)	0.63$_{\pm 0.01}$	0.31$_{\pm 0.03}$	0.10$_{\pm 0.03}$	0.47$_{\pm 0.02}$	−0.07$_{\pm 0.02}$	0.39$_{\pm 0.05}$	0.45$_{\pm 0.02}$	−0.04$_{\pm 0.02}$	0.39$_{\pm 0.05}$	0.61$_{\pm 0.02}$	0.06$_{\pm 0.02}$	0.24$_{\pm 0.02}$
	P(True) Sampling	0.63$_{\pm 0.01}$	0.24$_{\pm 0.03}$	0.10$_{\pm 0.03}$	0.46$_{\pm 0.02}$	−0.04$_{\pm 0.03}$	0.36$_{\pm 0.04}$	–	–	–	–	–	–
	P(True) Empirical	0.67$_{\pm 0.01}$	0.34$_{\pm 0.03}$	0.07$_{\pm 0.02}$	0.52$_{\pm 0.02}$	0.03$_{\pm 0.03}$	0.28$_{\pm 0.07}$	0.49$_{\pm 0.02}$	−0.01$_{\pm 0.02}$	0.38$_{\pm 0.05}$	0.33$_{\pm 0.02}$	−0.21$_{\pm 0.02}$	0.50$_{\pm 0.01}$
Black-box	NumSet	0.47$_{\pm 0.01}$	−0.05$_{\pm 0.03}$	0.38$_{\pm 0.06}$	0.70$_{\pm 0.02}$	0.33$_{\pm 0.03}$	0.08$_{\pm 0.02}$	0.55$_{\pm 0.01}$	0.08$_{\pm 0.02}$	0.24$_{\pm 0.06}$	0.45$_{\pm 0.01}$	−0.05$_{\pm 0.03}$	0.43$_{\pm 0.04}$
	LabelProb	0.51$_{\pm 0.01}$	0.05$_{\pm 0.03}$	0.27$_{\pm 0.05}$	0.66$_{\pm 0.02}$	0.30$_{\pm 0.03}$	0.14$_{\pm 0.04}$	0.59$_{\pm 0.02}$	0.19$_{\pm 0.02}$	0.08$_{\pm 0.03}$	0.45$_{\pm 0.02}$	−0.09$_{\pm 0.02}$	0.45$_{\pm 0.04}$
	Kernel Language Entropy (KLE)	0.49$_{\pm 0.01}$	0.01$_{\pm 0.03}$	0.33$_{\pm 0.05}$	0.72$_{\pm 0.02}$	0.39$_{\pm 0.03}$	0.10$_{\pm 0.03}$	0.64$_{\pm 0.02}$	0.29$_{\pm 0.02}$	0.06$_{\pm 0.02}$	0.40$_{\pm 0.02}$	−0.17$_{\pm 0.02}$	0.46$_{\pm 0.03}$
	EigValLap NLI (entail)	0.51$_{\pm 0.01}$	0.03$_{\pm 0.03}$	0.30$_{\pm 0.05}$	0.71$_{\pm 0.02}$	0.36$_{\pm 0.03}$	0.08$_{\pm 0.02}$	0.62$_{\pm 0.02}$	0.26$_{\pm 0.02}$	0.07$_{\pm 0.02}$	0.40$_{\pm 0.02}$	−0.18$_{\pm 0.02}$	0.47$_{\pm 0.02}$
	EigValLap NLI (contra)	0.42$_{\pm 0.01}$	−0.20$_{\pm 0.03}$	0.47$_{\pm 0.02}$	0.66$_{\pm 0.02}$	0.30$_{\pm 0.03}$	0.12$_{\pm 0.04}$	0.68$_{\pm 0.02}$	0.38$_{\pm 0.02}$	0.05$_{\pm 0.02}$	0.44$_{\pm 0.02}$	−0.10$_{\pm 0.02}$	0.45$_{\pm 0.03}$
	EigValLap Jaccard	0.42$_{\pm 0.01}$	−0.04$_{\pm 0.03}$	0.43$_{\pm 0.02}$	0.62$_{\pm 0.02}$	0.21$_{\pm 0.03}$	0.14$_{\pm 0.04}$	0.60$_{\pm 0.02}$	0.19$_{\pm 0.02}$	0.09$_{\pm 0.03}$	0.40$_{\pm 0.02}$	−0.13$_{\pm 0.02}$	0.49$_{\pm 0.02}$
	Eccentricity NLI (entail)	0.52$_{\pm 0.01}$	0.05$_{\pm 0.03}$	0.26$_{\pm 0.04}$	0.66$_{\pm 0.02}$	0.30$_{\pm 0.03}$	0.15$_{\pm 0.04}$	0.62$_{\pm 0.02}$	0.25$_{\pm 0.02}$	0.07$_{\pm 0.02}$	0.44$_{\pm 0.02}$	−0.13$_{\pm 0.02}$	0.42$_{\pm 0.04}$
	Eccentricity NLI (contra)	0.41$_{\pm 0.01}$	−0.22$_{\pm 0.03}$	0.47$_{\pm 0.03}$	0.71$_{\pm 0.02}$	0.37$_{\pm 0.03}$	0.10$_{\pm 0.03}$	0.67$_{\pm 0.02}$	0.36$_{\pm 0.02}$	0.06$_{\pm 0.02}$	0.44$_{\pm 0.02}$	−0.10$_{\pm 0.02}$	0.45$_{\pm 0.03}$
	Eccentricity Jaccard	0.57$_{\pm 0.01}$	0.20$_{\pm 0.03}$	0.26$_{\pm 0.05}$	0.60$_{\pm 0.02}$	0.19$_{\pm 0.03}$	0.14$_{\pm 0.04}$	0.49$_{\pm 0.02}$	−0.01$_{\pm 0.02}$	0.38$_{\pm 0.06}$	0.53$_{\pm 0.01}$	0.03$_{\pm 0.03}$	0.24$_{\pm 0.06}$
	DegMat NLI (entail)	0.51$_{\pm 0.01}$	0.05$_{\pm 0.03}$	0.29$_{\pm 0.05}$	0.72$_{\pm 0.02}$	0.37$_{\pm 0.03}$	0.08$_{\pm 0.03}$	0.62$_{\pm 0.02}$	0.25$_{\pm 0.02}$	0.06$_{\pm 0.02}$	0.40$_{\pm 0.02}$	−0.17$_{\pm 0.02}$	0.47$_{\pm 0.02}$
	DegMat NLI (contra)	0.39$_{\pm 0.01}$	−0.25$_{\pm 0.03}$	0.47$_{\pm 0.02}$	0.67$_{\pm 0.02}$	0.30$_{\pm 0.03}$	0.14$_{\pm 0.03}$	**0.68**$_{\pm 0.02}$	**0.38**$_{\pm 0.02}$	**0.05**$_{\pm 0.01}$	0.44$_{\pm 0.02}$	−0.10$_{\pm 0.02}$	0.45$_{\pm 0.04}$
	DegMat Jaccard	0.45$_{\pm 0.01}$	0.02$_{\pm 0.03}$	0.39$_{\pm 0.02}$	0.62$_{\pm 0.02}$	0.21$_{\pm 0.03}$	0.13$_{\pm 0.04}$	0.60$_{\pm 0.02}$	0.19$_{\pm 0.02}$	0.09$_{\pm 0.03}$	0.40$_{\pm 0.02}$	−0.14$_{\pm 0.02}$	0.49$_{\pm 0.02}$
	LUQ	0.41$_{\pm 0.01}$	−0.22$_{\pm 0.03}$	0.42$_{\pm 0.03}$	0.72$_{\pm 0.02}$	0.39$_{\pm 0.03}$	0.09$_{\pm 0.03}$	0.68$_{\pm 0.02}$	0.38$_{\pm 0.02}$	0.05$_{\pm 0.02}$	0.40$_{\pm 0.02}$	−0.16$_{\pm 0.02}$	0.46$_{\pm 0.03}$
	Lexical Sim. (ROUGE-L)	0.43$_{\pm 0.01}$	−0.01$_{\pm 0.03}$	0.43$_{\pm 0.02}$	0.65$_{\pm 0.02}$	0.25$_{\pm 0.03}$	0.11$_{\pm 0.03}$	0.55$_{\pm 0.02}$	0.12$_{\pm 0.02}$	0.15$_{\pm 0.03}$	0.34$_{\pm 0.02}$	−0.21$_{\pm 0.02}$	0.49$_{\pm 0.01}$
	Lexical Sim. (BLEU)	0.46$_{\pm 0.01}$	0.02$_{\pm 0.03}$	0.39$_{\pm 0.03}$	0.61$_{\pm 0.02}$	0.19$_{\pm 0.03}$	0.16$_{\pm 0.04}$	0.58$_{\pm 0.02}$	0.16$_{\pm 0.02}$	0.11$_{\pm 0.03}$	0.32$_{\pm 0.02}$	−0.23$_{\pm 0.02}$	0.50$_{\pm 0.01}$
AlignScore	AS-Precision(gold_answer)	**0.74**$_{\pm 0.01}$	**0.54**$_{\pm 0.02}$	**0.06**$_{\pm 0.02}$	0.85$_{\pm 0.02}$	0.75$_{\pm 0.03}$	0.24$_{\pm 0.04}$	0.69$_{\pm 0.02}$	0.34$_{\pm 0.02}$	0.06$_{\pm 0.02}$	**0.74**$_{\pm 0.02}$	**0.51**$_{\pm 0.03}$	**0.14**$_{\pm 0.03}$
	AS-Precision(wikipage)	–	–	–	0.81$_{\pm 0.01}$	0.50$_{\pm 0.03}$	**0.06**$_{\pm 0.02}$	**0.71**$_{\pm 0.02}$	**0.39**$_{\pm 0.02}$	**0.04**$_{\pm 0.01}$	–	–	–
	AS-Recall(gold_answer)	0.50$_{\pm 0.01}$	0.12$_{\pm 0.03}$	0.32$_{\pm 0.03}$	**0.90**$_{\pm 0.01}$	**0.81**$_{\pm 0.02}$	0.18$_{\pm 0.04}$	0.61$_{\pm 0.02}$	0.16$_{\pm 0.02}$	0.13$_{\pm 0.03}$	0.62$_{\pm 0.02}$	0.45$_{\pm 0.03}$	0.38$_{\pm 0.02}$
	AS-Recall(wikipage)	–	–	–	0.50$_{\pm 0.02}$	0.01$_{\pm 0.03}$	0.33$_{\pm 0.05}$	0.57$_{\pm 0.02}$	0.10$_{\pm 0.02}$	0.15$_{\pm 0.04}$	–	–	–

D.2 Agreement between estimator rankings

The first plots compare estimator rankings across models within each task; the second compare rankings across tasks for each model.

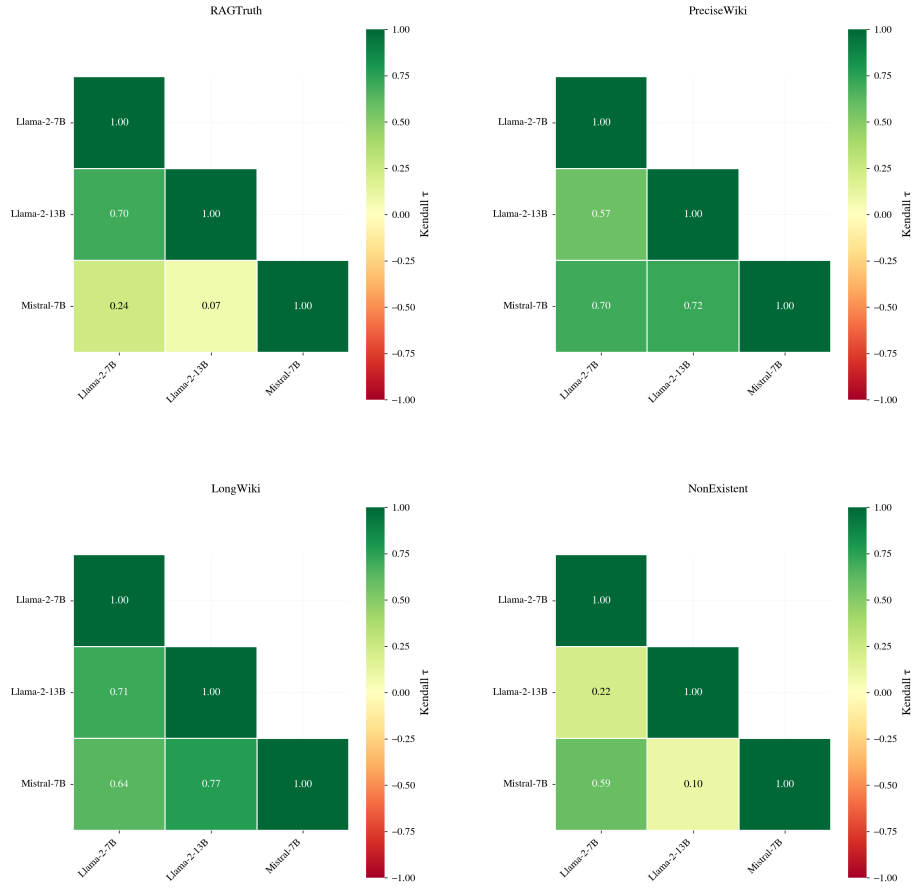


Figure 4: Pairwise Kendall's τ agreement between estimator rankings induced by ROC-AUC across models, shown separately for each task. Higher values indicate that the same estimators tend to rank similarly across models for a given hallucination type.

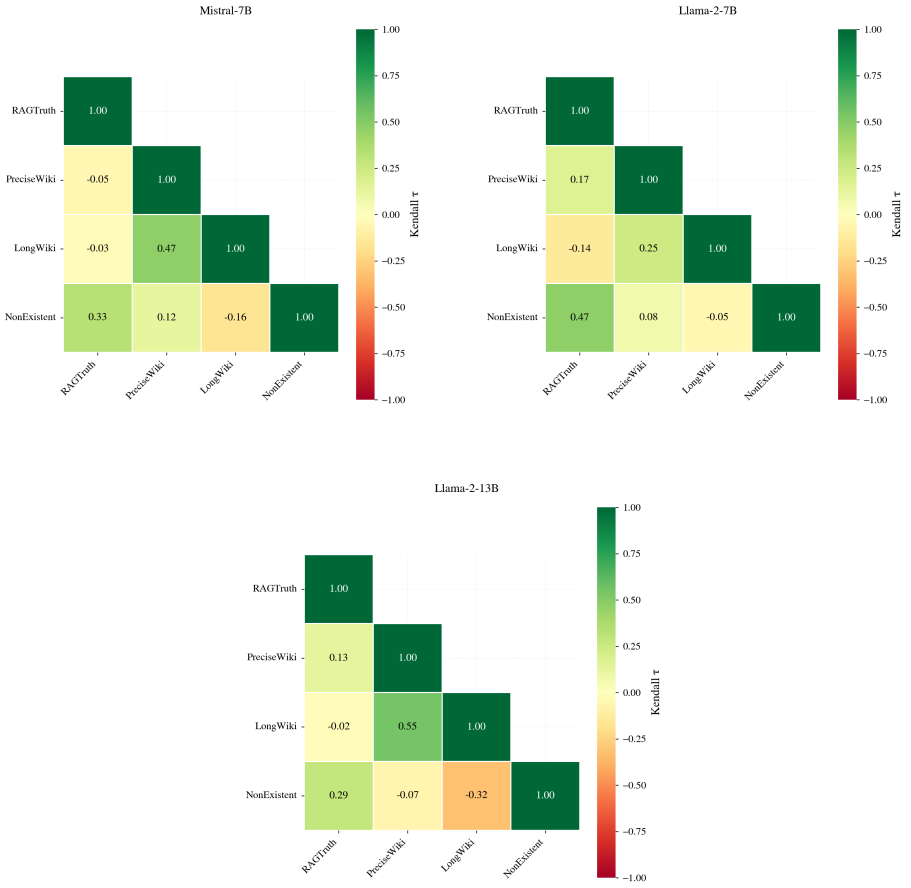


Figure 5: Pairwise Kendall's τ agreement between estimator rankings induced by ROC-AUC across tasks, shown separately for each model. Higher values indicate that the same estimators tend to rank similarly across hallucination types for a given model.

Table 10: Spearman correlation between estimator rankings induced by each pair of metrics, computed over uncertainty estimators. AlignScore variants are excluded. RCE is lower-is-better, so negative values indicate agreement. PRR is computed over the full rejection domain.

Task	Model	AUROC–PRR	AUROC–AUPRC	AUROC–RCE	PRR–RCE	AUPRC–RCE
RAGTruth	Llama 2 13B	.95	.99	-.96	-.91	-.95
–	Llama 2 7B	.91	.96	-.94	-.88	-.91
–	Mistral 7B	.95	.94	-.93	-.90	-.86
PreciseWikiQA	Llama 2 13B	.98	.98	-.92	-.91	-.94
–	Llama 2 7B	.96	.94	-.72	-.60	-.79
–	Mistral 7B	.97	.98	-.74	-.72	-.78
LongWiki	Llama 2 13B	.97	.99	-.92	-.94	-.92
–	Llama 2 7B	.95	.98	-.90	-.89	-.87
–	Mistral 7B	.94	.96	-.89	-.92	-.90
NonExistentRefusal	Llama 2 13B	.98	.94	-.92	-.91	-.83
–	Llama 2 7B	.94	.86	-.93	-.83	-.80
–	Mistral 7B	.98	.99	-.93	-.90	-.91

Agreement between evaluation metrics. This table holds the task and model fixed and compares rankings under different evaluation metrics.

D.3 Performance by task and estimator family

The rank table and per-task plots expand the main performance summary. The family curves complete the main-text comparison with LongWiki and NonExistentRefusal.

Table 11: Within each benchmark panel (one task and one model), estimators are ranked by the metric value (rank 1 is best for that metric). Each cell gives the mean of that rank across panels in the slice, with a subscript giving +/- one standard deviation of the rank across those panels. Lower mean rank is better. Bold marks the lowest mean rank in the column separately within uncertainty estimators and within AlignScore variants (same convention as the aggregate metrics table). Underline marks second-lowest mean rank among uncertainty estimators only; AlignScore rows are not underlined for second place. The integer in parentheses counts how many panels in the slice had rank at most K (here K=3), i.e. were in the top-K for that metric within that panel. Slice: all tasks and models pooled.

Estimator	AUROC	PRR	RCE
Output Token Count	$21.2_{\pm 17.3}$ (3)	$23.2_{\pm 20.1}$ (3)	$23.6_{\pm 17.7}$ (3)
Maximum Seq. Probability (MSP)	$13.5_{\pm 10.2}$ (1)	$14.9_{\pm 12.3}$ (1)	$14.8_{\pm 10.2}$ (2)
Perplexity (PPL)	$28.2_{\pm 10.1}$ (0)	$29.7_{\pm 8.8}$ (0)	$26.1_{\pm 12.9}$ (0)
Mean Token Entropy (MTE)	$24.4_{\pm 9.9}$ (0)	$25.7_{\pm 10.5}$ (0)	$23.0_{\pm 10.0}$ (0)
Pointwise Mutual Info. (PMI)	$40.8_{\pm 7.8}$ (0)	$39.4_{\pm 6.4}$ (0)	$37.2_{\pm 10.2}$ (0)
Conditional PMI	$33.6_{\pm 18.1}$ (0)	$32.7_{\pm 19.2}$ (0)	$33.9_{\pm 18.3}$ (1)
Self Certainty	$29.1_{\pm 9.1}$ (0)	$27.3_{\pm 11.0}$ (0)	$25.6_{\pm 11.5}$ (0)
Rényi Divergence	$31.2_{\pm 9.0}$ (0)	$29.1_{\pm 10.6}$ (0)	$27.8_{\pm 12.2}$ (0)
Fisher–Rao Distance	$27.8_{\pm 11.3}$ (0)	$26.4_{\pm 12.6}$ (0)	$25.8_{\pm 10.2}$ (0)
Token SAR	$25.0_{\pm 11.4}$ (0)	$25.8_{\pm 10.1}$ (0)	$24.6_{\pm 12.1}$ (0)
Claim-Conditioned Probability (CCP)	**$8.7_{\pm 5.7}$ (4)**	$\underline{10.9_{\pm 7.4}\ (3)}$	**$10.2_{\pm 7.6}$ (3)**
Monte-Carlo Seq. Entropy	$25.3_{\pm 10.0}$ (0)	$24.0_{\pm 10.1}$ (0)	$22.0_{\pm 11.7}$ (1)
Monte-Carlo Norm. Seq. Entropy	$32.0_{\pm 7.8}$ (0)	$30.0_{\pm 6.3}$ (0)	$29.4_{\pm 12.3}$ (0)
Semantic Entropy	$23.6_{\pm 9.4}$ (0)	$21.9_{\pm 8.3}$ (0)	$23.0_{\pm 13.1}$ (2)
Semantic Density	$20.0_{\pm 11.5}$ (0)	$19.4_{\pm 11.5}$ (0)	$20.6_{\pm 11.8}$ (1)
Sentence SAR	$17.4_{\pm 14.5}$ (2)	$17.8_{\pm 14.3}$ (1)	$19.7_{\pm 14.3}$ (1)
SAR	$22.8_{\pm 9.9}$ (0)	$22.9_{\pm 10.3}$ (0)	$25.7_{\pm 11.7}$ (0)
Cocoa MSP	$\underline{10.3_{\pm 5.2}\ (1)}$	**$10.8_{\pm 4.5}$ (1)**	$\underline{14.2_{\pm 9.4}\ (0)}$
Cocoa PPL	$24.3_{\pm 13.3}$ (0)	$24.1_{\pm 12.2}$ (0)	$25.3_{\pm 13.3}$ (1)
Cocoa MTE	$21.2_{\pm 11.5}$ (1)	$21.2_{\pm 11.8}$ (0)	$21.2_{\pm 12.5}$ (1)
Attention Score	$22.2_{\pm 19.4}$ (3)	$22.3_{\pm 19.8}$ (5)	$23.0_{\pm 18.6}$ (2)
RAUQ	$25.8_{\pm 10.4}$ (0)	$25.6_{\pm 9.7}$ (0)	$25.9_{\pm 13.4}$ (1)
CSL	$33.5_{\pm 7.5}$ (0)	$32.2_{\pm 8.4}$ (0)	$32.5_{\pm 7.5}$ (0)
EigenScore	$29.7_{\pm 14.1}$ (0)	$27.3_{\pm 13.1}$ (0)	$31.5_{\pm 11.8}$ (0)
Mahalanobis Distance (MD)	$45.0_{\pm 10.4}$ (0)	$44.4_{\pm 10.8}$ (0)	$40.9_{\pm 10.2}$ (0)
Relative MD (RMD)	$44.8_{\pm 9.1}$ (0)	$43.5_{\pm 9.3}$ (0)	$42.6_{\pm 10.0}$ (0)
Robust Density Estimation (RDE)	$37.1_{\pm 13.4}$ (0)	$38.7_{\pm 11.9}$ (0)	$35.1_{\pm 11.9}$ (0)
HUQ-MD	$28.8_{\pm 10.0}$ (0)	$30.0_{\pm 9.7}$ (0)	$25.9_{\pm 13.8}$ (1)
HUQ-RMD	$29.0_{\pm 9.4}$ (0)	$30.1_{\pm 8.7}$ (0)	$26.2_{\pm 13.5}$ (1)
P(True)	$35.5_{\pm 15.0}$ (0)	$35.7_{\pm 13.3}$ (0)	$33.7_{\pm 13.5}$ (0)
P(True) Sampling	$31.2_{\pm 16.2}$ (0)	$31.3_{\pm 14.3}$ (0)	$30.7_{\pm 16.9}$ (0)
P(True) Empirical	$31.2_{\pm 15.2}$ (0)	$31.8_{\pm 14.8}$ (0)	$31.3_{\pm 15.3}$ (1)
NumSet	$23.9_{\pm 10.5}$ (0)	$28.8_{\pm 11.5}$ (0)	$27.2_{\pm 11.2}$ (0)
LabelProb	$22.8_{\pm 9.4}$ (0)	$22.6_{\pm 9.1}$ (0)	$27.6_{\pm 10.3}$ (0)
Kernel Language Entropy (KLE)	$20.5_{\pm 11.0}$ (0)	$19.9_{\pm 12.6}$ (0)	$22.5_{\pm 12.6}$ (1)
EigValLap NLI (entail)	$21.1_{\pm 11.2}$ (1)	$21.4_{\pm 12.5}$ (0)	$21.5_{\pm 9.7}$ (0)
EigValLap NLI (contra)	$19.3_{\pm 12.1}$ (2)	$22.3_{\pm 14.0}$ (2)	$21.0_{\pm 14.0}$ (0)
EigValLap Jaccard	$31.3_{\pm 10.1}$ (0)	$29.5_{\pm 10.6}$ (0)	$28.2_{\pm 14.9}$ (0)
Eccentricity NLI (entail)	$19.8_{\pm 9.5}$ (0)	$20.6_{\pm 10.1}$ (0)	$21.2_{\pm 8.7}$ (0)
Eccentricity NLI (contra)	$17.6_{\pm 11.9}$ (1)	$20.8_{\pm 14.5}$ (1)	$18.5_{\pm 13.4}$ (0)
Eccentricity Jaccard	$24.7_{\pm 12.1}$ (0)	$25.5_{\pm 13.6}$ (0)	$26.7_{\pm 13.4}$ (0)
DegMat NLI (entail)	$21.5_{\pm 12.7}$ (0)	$21.6_{\pm 12.7}$ (0)	$18.2_{\pm 12.6}$ (1)
DegMat NLI (contra)	$21.3_{\pm 13.0}$ (1)	$23.9_{\pm 14.9}$ (1)	$22.2_{\pm 15.8}$ (2)
DegMat Jaccard	$29.3_{\pm 9.6}$ (0)	$26.6_{\pm 9.5}$ (0)	$26.5_{\pm 13.0}$ (0)
LUQ	$21.2_{\pm 14.9}$ (0)	$22.1_{\pm 16.5}$ (1)	$23.6_{\pm 13.0}$ (1)
Lexical Sim. (ROUGE-L)	$33.7_{\pm 8.2}$ (0)	$31.5_{\pm 8.0}$ (0)	$28.5_{\pm 12.9}$ (0)
Lexical Sim. (BLEU)	$31.6_{\pm 9.0}$ (0)	$29.3_{\pm 9.5}$ (0)	$25.8_{\pm 13.2}$ (1)
AS-Precision(gold_answer)	$5.7_{\pm 6.1}$ (6)	**$3.5_{\pm 2.6}$ (8)**	$12.2_{\pm 13.2}$ (3)
AS-Precision(wikipage)	**$2.7_{\pm 1.7}$ (5)**	$4.8_{\pm 5.3}$ (4)	**$8.4_{\pm 8.7}$ (3)**
AS-Recall(gold_answer)	$4.7_{\pm 7.3}$ (5)	$7.0_{\pm 6.2}$ (5)	$20.9_{\pm 11.5}$ (0)
AS-Recall(wikipage)	$35.0_{\pm 11.1}$ (0)	$33.5_{\pm 9.5}$ (0)	$32.8_{\pm 11.7}$ (0)

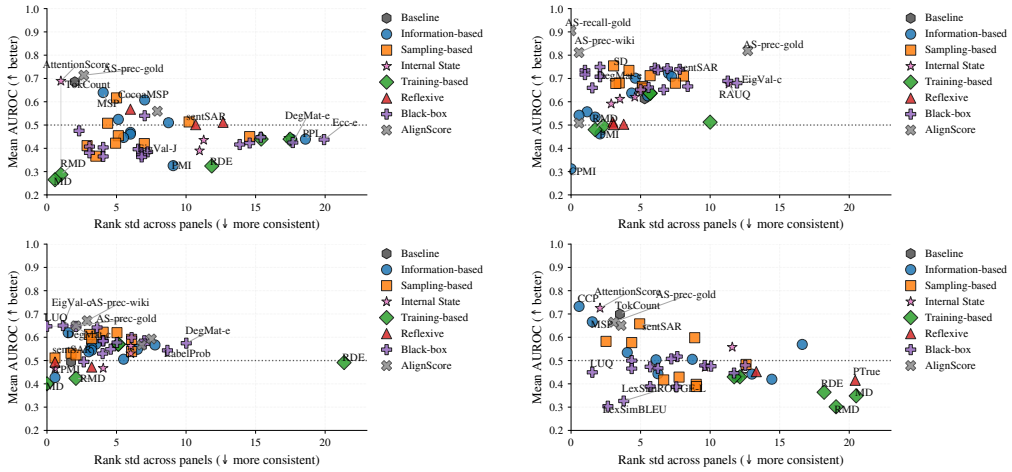


Figure 6: Performance–stability profiles for uncertainty estimators within each task, aggregated across the three models. Each point represents one estimator or AlignScore variant. The vertical axis reports mean ROC-AUC across models for the corresponding task, while the horizontal axis reports the standard deviation of the estimator's within-model rank; lower values indicate more stable rankings across models. The panels show, from left to right and top to bottom: RAGTruth , PreciseWikiQA , LongWiki , and NonExistentRefusal . AlignScore variants are reference-based comparators rather than uncertainty estimators.

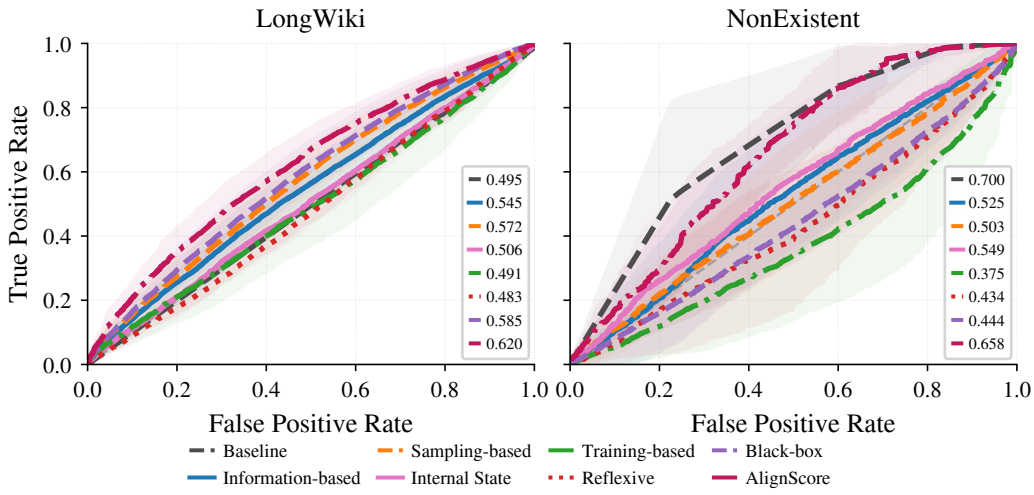


Figure 7: Family-level ROC aggregates per task, averaged across the three models. For each task–model panel, estimator ROC curves are first averaged within each family by taking the mean true positive rate at each false positive rate. For each task, these model-level family means are then averaged across models. Shaded bands show one standard deviation of these model-level family-mean true positive rates at each false positive rate; they reflect cross-model variability.

D.4 Response length

For the length comparison in Section 5, we compare estimator AUROC with token count on the same responses, then evaluate performance within five length groups.

For paired AUROC differences, both scores use the same 1,000 resampled sets of responses. Intervals are the 2.5th and 97.5th percentiles, with seed 42. The grouping identifier is exact input text because a source-document identifier is unavailable in the aligned inputs. All retained inputs are unique within each panel, so this reduces to response-level resampling. Dependence between different prompts from one source is not accounted for. All resamples in the reported RAGTruth comparisons contained both classes. The intervals describe each comparison separately; they do not establish simultaneous differences across all methods.

RAGTruth uses five approximately equal length groups without splitting tied token counts; both classes occur in every group. LongWiki and NonExistentRefusal responses concentrate at generation limits, preventing a comparable check. For example, 99.45% of valid Llama 2 13B LongWiki responses reach the 256-token cap.

These groups retain some length variation and do not isolate a causal effect of length.

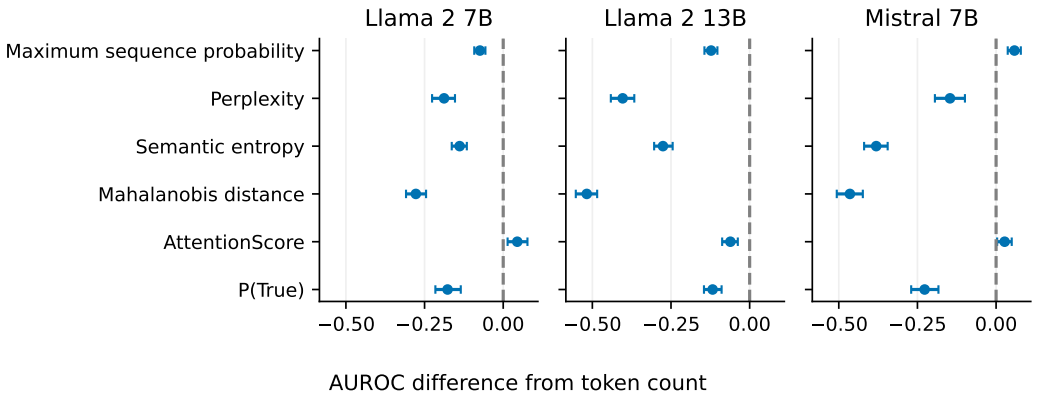


Figure 8: RAGTruth AUROC difference from token count. Points to the right of zero favor the estimator. Bars are paired 95% bootstrap intervals (1,000 resamples), calculated separately for each comparison.

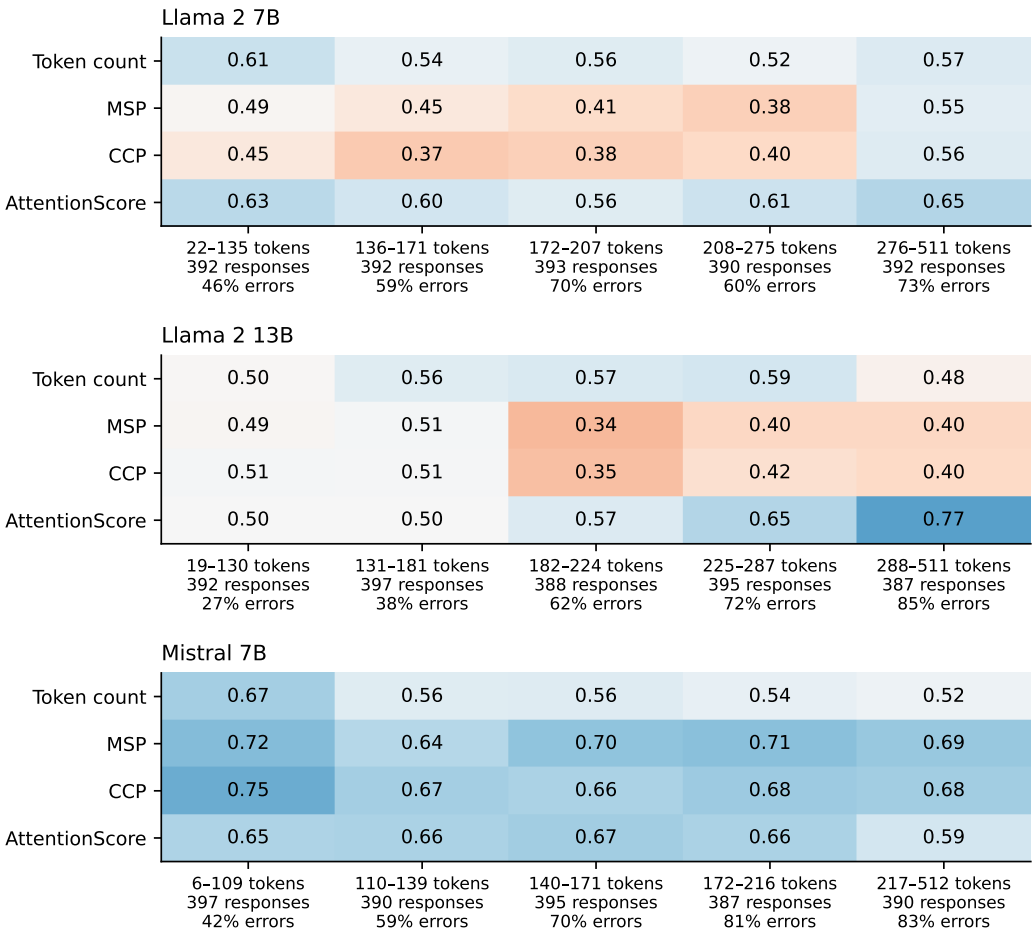


Figure 9: RAGTruth AUROC within response-length groups, for the signals discussed in Section 5. Columns run from shortest to longest responses and give the token range, sample size and percentage of errors. AUROC 0.5 is chance; larger values are better. MSP: maximum sequence probability; CCP: claim-conditioned probability. Tied lengths stay together; both classes occur in every group.

D.5 CHANGING THE JUDGE

The judge comparison in Section 5 checks whether relative estimator performance depends on the primary HalluLens evaluator. We replace the Llama judge with Qwen2.5-72B-Instruct-AWQ while keeping generated responses and uncertainty scores fixed. On PreciseWikiQA, judge-specific exclusions can change the evaluated subset, so differences reflect both judgments and response selection. On LongWiki, extracted claims and retrieved evidence also remain fixed; only claim verification changes.

Judge replacement measures sensitivity, not label correctness. The quantized judge used here has not been independently validated against human judgments.

The nine task–model comparisons summarized in Section 5 have AUROC rank correlations from 0.906 to 0.995 under judge replacement. For LongWiki, agreement on continuous quality scores is weaker: Spearman correlations are 0.936, 0.930 and 0.431 for Llama 2 7B, Llama 2 13B and Mistral 7B. Mistral's binary labels agree on 71.86% of responses. Thus, similar estimator rankings need not mean similar judgments of individual responses.

Table 12: PreciseWikiQA response selection under the two judges. Both evaluate the same 2,000 generated responses per model; each excludes its own abstentions and unverifiable judgments. Errors are incorrect answers among the retained responses.

Model	Judge	Evaluated	Errors	Excluded
Llama 2 7B	Llama	1,767	1,637	233
Llama 2 7B	Qwen	1,489	1,334	511
Llama 2 13B	Llama	1,923	1,672	77
Llama 2 13B	Qwen	1,718	1,443	282
Mistral 7B	Llama	1,955	1,686	45
Mistral 7B	Qwen	1,870	1,596	130

D.6 RAGTruth: error types and proportions

The comparisons in Section 5 use the same 1,959 responses and 45 estimators per model, changing only the definition of an error.

For RAGTruth, the four binary targets indicate any annotated span, conflict with the context, baseless information unsupported by it, and the presence of both subtypes. Conflict includes the evident and subtle annotations grouped by the processed dataset. Subtypes are inclusive: a response with both is positive for each subtype. We also measure prevalence within each response as the proportion of tokens covered by the union of annotated spans. Overlapping spans are counted once. This is distinct from the fraction of responses labeled as errors. The continuous quality target is one minus this token proportion.

The distribution plot separates responses with one subtype from those with both; the evaluation targets include co-occurring subtypes as described above. PRR@50 restricts rejection to half the responses,

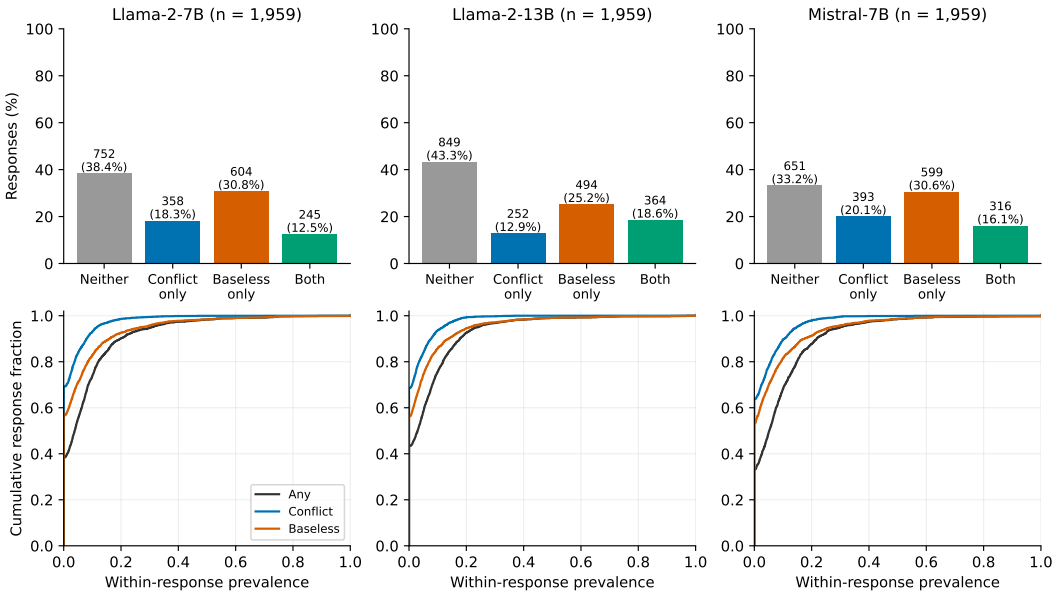


Figure 10: RAGTruth subtype distributions for each model. Top: mutually exclusive occurrence categories with counts and percentages of all 1,959 responses. Bottom: empirical cumulative distributions of within-response prevalence, defined as the fraction of response tokens covered by annotated spans. All responses are included; the jump at zero represents responses without that annotation. Any combines both subtypes; overlapping spans are counted once. These distributions describe the targets rather than estimator performance.

as a sensitivity check on the integration range. Main results use the full range. Ranking estimators by direct score–target correlation gives positive agreement between subtypes, so the reversal does not apply to every evaluation criterion.

Changing the minimum annotated proportion. The following check requires an annotated span of any length, or at least 10% or 25% of response tokens to be annotated. This changes the binary error labels used for AUROC; the continuous targets used for PRR and RCE remain unchanged. The thresholds are sensitivity checks, not selected operating points.

Table 13: Best observed metric values across the 45 estimators for each occurrence target. Each row contains 1,959 responses; Errors counts positive labels and Error rate gives their fraction (the constant-score AUPRC baseline). Maxima are reported for AUROC, AUPRC and PRR; minima for RCE. The best value in each column can come from a different estimator and is not a validated selection rule.

Model	Target	Errors	Error rate	AUROC	AUPRC	PRR	RCE
Llama 2 7B	Any	1207	0.616	0.647	0.701	0.311	0.148
Llama 2 7B	Conflict	603	0.308	0.678	0.426	0.456	0.093
Llama 2 7B	Baseless	849	0.433	0.584	0.516	0.224	0.132
Llama 2 7B	Both	245	0.125	0.689	0.192	0.538	0.093
Llama 2 13B	Any	1110	0.567	0.690	0.699	0.388	0.070
Llama 2 13B	Conflict	616	0.314	0.751	0.515	0.602	0.070
Llama 2 13B	Baseless	858	0.438	0.673	0.558	0.449	0.040
Llama 2 13B	Both	364	0.186	0.771	0.353	0.692	0.055
Mistral 7B	Any	1308	0.668	0.761	0.842	0.560	0.033
Mistral 7B	Conflict	709	0.362	0.724	0.544	0.570	0.025
Mistral 7B	Baseless	915	0.467	0.684	0.629	0.464	0.025
Mistral 7B	Both	316	0.161	0.767	0.344	0.665	0.040

Table 14: Spearman agreement between conflict and baseless estimator rankings, using the same 1,959 responses and 45 estimators per model. PRR integrates over the full rejection range; PRR@50 permits rejection of at most half the responses. RCE uses native quality. The last row in each block ranks estimators by their Spearman association with the binary error or token-share target; it is not a direct correlation between the two targets. For prevalence, the top-five PRR sets have no overlap for either rejection range.

Target	Evaluation	Llama 2 7B	Llama 2 13B	Mistral 7B
Occurrence	AUROC	0.613	0.846	0.933
Occurrence	AUPRC	0.519	0.857	0.888
Occurrence	PRR	0.679	0.821	0.921
Occurrence	PRR@50	0.546	0.849	0.901
Occurrence	RCE	0.511	0.878	0.912
Occurrence	Score–target ρ	0.591	0.843	0.933
Prevalence	PRR	-0.251	-0.123	-0.561
Prevalence	PRR@50	-0.425	-0.544	-0.604
Prevalence	RCE	-0.514	-0.222	-0.528
Prevalence	Score–target ρ	0.179	0.591	0.716

Table 15: RAGTruth error counts out of 1,959 responses per model for the thresholds in Figure 11. Remaining responses are non-errors. Percentages in the threshold column refer to the share of annotated response tokens, not the fraction of erroneous responses.

Annotation	Token share	Llama 2 7B	Llama 2 13B	Mistral 7B
Any annotated tokens	Any amount	1207	1110	1308
Any annotated tokens	$\geq 10\%$	516	479	624
Any annotated tokens	$\geq 25\%$	134	90	150
Baseless information	Any amount	849	858	915
Baseless information	$\geq 10\%$	338	279	364
Baseless information	$\geq 25\%$	110	74	114
Conflict with context	Any amount	603	616	709
Conflict with context	$\geq 10\%$	137	126	201

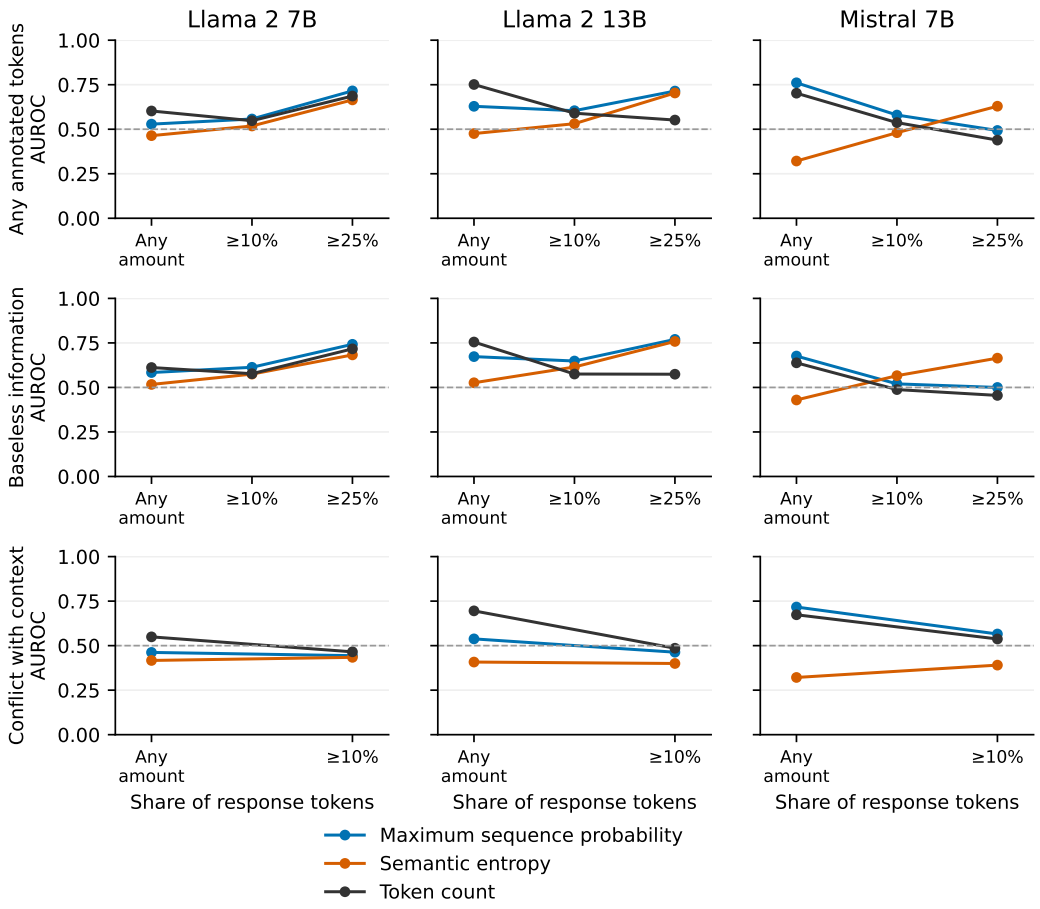


Figure 11: RAGTruth AUROC for three fixed examples as the error definition changes. “Any amount” means at least one annotated span of the indicated type; 10% and 25% require that share of response tokens to be annotated. Rows distinguish all annotations, baseless information, and conflict. The dashed line marks chance (0.5). Lines connect computed settings only; conflict was not evaluated at 25%. Class counts are in Table 15.

D.7 LongWiki: precision, recall and error thresholds

F1@32 combines claim precision with a fixed quota of supported claims (Section C.1). For example, ten supported claims and no unsupported claims give precision 1 but F1@32 of 0.476. We keep responses and uncertainty scores fixed and compare the labels and estimator rankings obtained from precision and Recall@32 separately. Cohen's κ measures label agreement after accounting for class frequencies; Spearman's ρ measures agreement between AUROC rankings.

Table 16: LongWiki precision versus Recall@32 at threshold 0.5. Errors have quality below 0.5. Label agreement is Cohen's κ; rank agreement is Spearman's ρ between AUROC rankings of 45 estimators on the same responses. The two agreement columns concern different objects: response labels and estimator rankings.

Model	Responses	Errors		Label agreement	Rank agreement
		Precision	Recall@32		
Llama 2 7B	1996	1506	1336	0.588	0.903
Llama 2 13B	1999	1262	1879	0.192	0.961
Mistral 7B	1997	1837	1617	0.457	0.844

Changing the F1@32 threshold. The tables below vary the cutoff used to label a response as an error for AUROC. Counts are reported at every audited threshold; performance is shown only where computed. PRR and RCE retain the continuous F1@32 target. This checks threshold sensitivity without selecting a preferred cutoff.

Table 17: LongWiki error / non-error counts under the strict rule F1@32 $< t$. Responses without an F1 target are excluded. These are target counts, not estimator performance estimates. At $t = 0.9$, AUROC is undefined for Llama 2 13B and Mistral 7B because only one class remains.

Threshold	Llama 2 7B	Llama 2 13B	Mistral 7B
0.1	124 / 1,872	202 / 1,797	216 / 1,781
0.3	747 / 1,249	993 / 1,006	1,128 / 869
0.5	1,480 / 516	1,729 / 270	1,785 / 212
0.7	1,869 / 127	1,973 / 26	1,967 / 30
0.9	1,992 / 4	1,999 / 0	1,997 / 0

Table 18: LongWiki AUROC as the F1@32 error threshold changes. A response is an error when F1@32 is strictly below the threshold. These three fixed examples illustrate threshold sensitivity; they are not selected winners. Class counts are in Table 17.

Model	Threshold	MSP	Semantic entropy	Token count
Llama 2 7B	0.3	0.540	0.506	0.469
Llama 2 7B	0.5	0.559	0.509	0.483
Llama 2 7B	0.7	0.660	0.582	0.478
Llama 2 13B	0.3	0.562	0.558	0.499
Llama 2 13B	0.5	0.589	0.577	0.499
Llama 2 13B	0.7	0.566	0.678	0.497
Mistral 7B	0.3	0.513	0.510	0.483
Mistral 7B	0.5	0.522	0.530	0.502
Mistral 7B	0.7	0.466	0.421	0.520

MSP: maximum sequence probability.