

The Role of Ambiguity in Error Prediction via Uncertainty Quantification

Ieva Raminta Staliūnaite^{✉, 📧} James Bishop[📧] Andreas Vlachos[✉]
[✉]University of Cambridge [📧]The Alan Turing Institute
{irs38, av308}@cam.ac.uk jbishop@turing.ac.uk

Abstract

The task of Error Prediction, namely predicting whether a model output is correct, is commonly tackled with Uncertainty Quantification (UQ). However, while uncertainty metrics capture when models lack knowledge or capacity to make a prediction, they also reflect aleatoric uncertainty, inherent in the input and context. We improve error prediction for Large Language Models (LLMs) by disentangling input ambiguity from the UQ signal. We experiment on Question Answering (QA) with six UQ metrics and show that UQ metrics are more predictive of errors on unambiguous instances than on questions with multiple plausible answers. We use Gated Experts and Selective Prediction to incorporate gold and predicted ambiguity labels into the error prediction pipeline. We find that ambiguity information improves error prediction across model families, training and evaluation paradigms, and datasets (including allegedly unambiguous ones), with the gains concentrated on the UQ metrics that most entangle aleatoric and epistemic uncertainty, gaining over 10 PRR points for individual metrics.

1 Introduction

Predicting whether a model will correctly answer a question is important for establishing model reliability and reducing risks associated with incorrect model outputs. The task of error prediction is usually performed by leveraging model uncertainty, yielding successful results (Kadavath et al., 2022; Lin et al., 2023; Manakul et al., 2023; Farquhar et al., 2024). However, uncertainty comprises both aleatoric and epistemic uncertainty. Aleatoric uncertainty is inherent to the data due to ambiguity, conflicting inputs or noise, and is therefore irreducible, whereas epistemic uncertainty, which stems from the model’s lack of capacity or knowledge, is reducible (Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021).

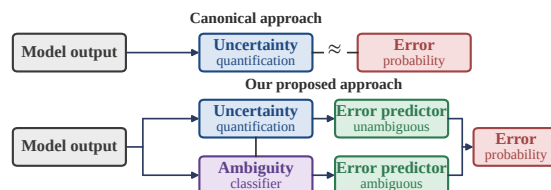


Figure 1: The established error prediction method directly using model features such as uncertainty metrics (top) and our proposed framework, predicting ambiguity and using it in a gated experts model for improving error prediction scores (bottom).

Ambiguity may be present at every step of the LLM pipeline. It can arise from conflicting information in the training data, a user asking an underspecified question, or contradictions in the context. Question Answering (QA) offers an experimental setup where ambiguity occurs naturally. To illustrate this, consider the following example from AmbigQA (Min et al., 2020): “What is it called when you mix up the letters of a word?” Valid answers include ‘dyslexia’ (referring to a learning disability) or ‘anagram’ (referring to an orthographic phenomenon). A model processing this question may be legitimately uncertain about the answer, given that there are at least two plausible interpretations. In such a case the canonical approach to error prediction (see top of Figure 1) would read the high model uncertainty as a high probability of error, since it does not discriminate aleatoric from epistemic uncertainty. As this uncertainty reflects input ambiguity rather than model error, the score is uninformative here rather than necessarily incorrect.

Some work has shown that UQ metrics capture both aleatoric and epistemic uncertainty, and the entanglement of the two makes UQ metrics less reliable for error prediction in ambiguous settings. Baan et al. (2024) posit that standard predictive distributions inherently conflate epistemic uncer-

tainty reflecting a model’s lack of knowledge, and aleatoric uncertainty reflecting inherent input ambiguity and valid human label variation. Tomov et al. (2025) formally prove that consistency-based and ensemble-based UQ estimators are guaranteed to correlate with epistemic uncertainty only when aleatoric uncertainty is zero, and show empirically that their performance degrades to near-random on ambiguous QA datasets.

In this work we focus on improving error prediction by leveraging ambiguity. We implement two approaches for separating out the aleatoric uncertainty in UQ metrics. First, a gated expert model uses the ambiguity signal as a routing mechanism, directing queries to separate, specialized sub-networks, one trained for ambiguous instances and one for unambiguous instances, to optimize error prediction for each specific domain independently. Second, a selective prediction model uses the ambiguity signal to explicitly adjust the uncertainty score, selectively rejecting error-prone answers by penalizing or favoring samples based on their ambiguity.

In our experiments we consider six Uncertainty Quantification (UQ) metrics, namely Maximum Softmax Probability (MSP) (Hendrycks and Gimpel, 2016), Semantic Entropy (Farquhar et al., 2024), Mutual Information via Iterative Prompting (MI) (Yadkori et al., 2024), Shifting Attention to Relevance (SAR) (Duan et al., 2024), Semantic Energy (Ma et al., 2025) and CoCoA (Vashurin et al., 2025). We carry out experiments on three datasets: AmbigQA (Min et al., 2020), NATCON-FQA (NCQA) (Nachshoni et al., 2025), and TriviaQA (Joshi et al., 2017). We show that ambiguity helps predict errors on both ambiguous and unambiguous QA datasets and across model families, including models finetuned to represent human label variation, with the largest gains on the uncertainty metrics that most entangle aleatoric and epistemic uncertainty and on their combination.¹

2 Related Work

Research has shown that UQ metrics encode more than whether the model is likely to be wrong. This distinction has been demonstrated to be relevant in Hate Speech Detection and Sentiment Analysis (Mostafazadeh Davani et al., 2022), Word Sense Disambiguation (Liu and Liu, 2023), Natural Lan-

guage Inference (Staliūnaitė and Vlachos, 2025), and Machine Translation (Staliūnaitė et al., 2026). However, the question of whether explicitly modeling ambiguity can improve error prediction, rather than task accuracy, remains open. Most related to our work, Cole et al. (2023) show that sampling-based confidence scores are better calibrated than likelihood on ambiguous questions, and propose a disambiguate-then-answer paradigm to handle denotational uncertainty. They also attempt to predict ambiguity from the model’s own outputs but report that none of their methods exceed chance by a meaningful margin, and consequently do not incorporate predicted ambiguity as an explicit signal in the abstention decision. Hou et al. (2024) and Walha et al. (2026) similarly decompose predictive uncertainty into aleatoric and epistemic components via input clarification ensembling and a spectral kernel-based approach respectively, but target uncertainty estimation itself rather than using the decomposed signal to supervise error prediction. Cox et al. (2025) instead address miscalibration under meaning-preserving paraphrases, a different source of variation from the answer multiplicity we model, while hybrid estimators such as HUQ (Vazhentsev et al., 2023) combine the two uncertainty types for selective classification over a fixed label set, which does not transfer directly to open-ended generation.

LLMs have been shown to be overconfident in their most likely prediction (Tian et al., 2023), and often confident in incorrect answers (Simhi et al., 2025). This is particularly relevant when multiple answers are equally plausible. Some work has shown that modeling ambiguity or subjectivity boosts task performance (Uma et al., 2021; Plank, 2022). Recent finetuning methods address this more directly, namely Sorensen et al. (2025) introduce Spectrum Tuning to enhance distributional coverage, and Zhang et al. (2025a) show that training on distributions or multi-sample prompting mitigates mode collapse. Furthermore, Yang et al. (2025) demonstrate that while the presence of inherent data uncertainty in multi-answer QA setups significantly degrades standard UQ metrics like max logit or verbalized confidence, entropy and response consistency remain robust estimators of model uncertainty. Yadkori et al. (2024) propose a UQ metric based on iterative prompting that explicitly targets epistemic uncertainty by measuring sensitivity to prior context. It requires multiple sequential prompting rounds and assumes aleatoric un-

¹The code is available on https://github.com/ieva-raminta/ambiguity_uncertainty_error

certainty manifests as context-insensitivity. None of these methods ask whether a model that better represents distributional uncertainty also becomes easier to supervise for error prediction.

Prior work has shown that internal model representations encode strong signals of output reliability and ambiguity. Namely, [Chen et al. \(2024\)](#) leverage eigenvalues of response embedding covariance to detect hallucinations, [Orgad et al. \(2024\)](#) show that probing exact-answer tokens reveals fine-grained error-type information, and [Li et al. \(2023\)](#) identify and steer truthfulness-related attention heads at inference time. Furthermore, [Vashurin et al. \(2025\)](#) train a lightweight model on middle-layer embeddings to predict consistency-based uncertainty scores without repeated sampling. [Ah-dritz et al. \(2024\)](#) train linear probes on internal embeddings to predict token-level confidence, finding middle layers most predictive. [CH-Wang et al. \(2024\)](#) show hallucination propensity can be detected from hidden states alone. We note that [Zhang et al. \(2025b\)](#) address a different task of classifying original questions against explicitly disambiguated counterparts, introducing systematic surface differences. Their signal peaks in early lexically-focused layers ([Jin et al., 2025](#)), consistent with probes capturing surface form rather than semantic ambiguity. This motivates including middle-layer representations as ambiguity classification features, where the signal of interest is semantic rather than surface-level.

Finally, [Huang et al. \(2026\)](#) show that many TriviaQA questions suffer from underspecification and that detecting and rewriting them improves answering accuracy. [Zhang et al. \(2025a\)](#) similarly show that standard single-answer benchmarks admit multiple valid responses, and introduce a reinforcement learning framework mining alternative answers to improve model performance. It is unknown whether the latent ambiguity in such datasets also undermines the ability of UQ metrics to predict errors on them.

3 Method

We extend the standard approach of using UQ metrics directly as error predictors, which we define formally as our baseline in Section 4.4 and compare against throughout. The injection of ambiguity labels is motivated by Figure 2, which illustrates that when there is little ambiguity (left hand-side of the figure), low uncertainty indicates low error

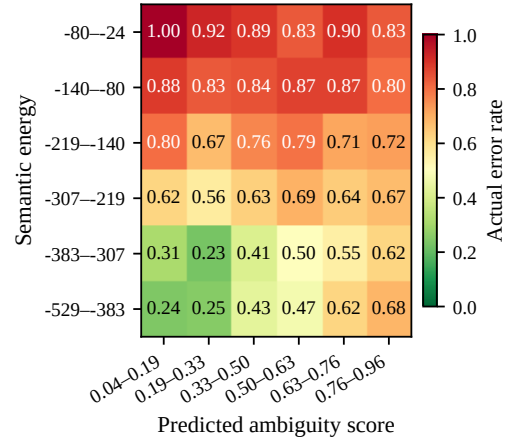


Figure 2: Heatmap illustrating the relationship between UQ (Semantic Energy) and predicted ambiguity on the one hand, and the error rate on the other. Ambiguity scores are predicted by the model described in Section 3.2. The same pattern holds with gold ambiguity labels (the annotated ambiguity labels of Section 4.1), confirming it reflects a genuine data property rather than a modelling artefact.

likelihood and high uncertainty indicates high error rates, whereas when ambiguity is high (right) the error likelihood is similar for low and high uncertainty scores. Conditioning on ambiguity thus lets the model trust UQ where it is informative and discount it where it is not.

3.1 Features for Error Prediction and Ambiguity Classification

We use three types of features to capture aleatoric and epistemic uncertainty signal, namely internal model representations, uncertainty quantification metrics, and semantic cluster statistics.

3.1.1 Uncertainty Quantification

We use UQ metrics as features in the ambiguity classifier (Section 5.2) and the error prediction models (Section 3.2). We consider six methods, computed over a greedy decode and K stochastic samples, that differ in how they aggregate confidence across samples.

MSP ([Hendrycks and Gimpel, 2016](#)) uses one minus the highest sequence-level probability across candidates as a simple confidence baseline. **Semantic Entropy** ([Farquhar et al., 2024](#)) clusters samples by bidirectional entailment and computes entropy over cluster probabilities, abstracting away lexical variation. **Semantic Energy** ([Ma et al., 2025](#)) replaces the cluster probability with a Boltzmann-style energy distribution over unnor-

malised logits, addressing cases where identical incorrect responses yield zero entropy. **CoCoA** (Vashurin et al., 2025) multiplies the greedy output’s confidence by its mean pairwise dissimilarity to the samples, retaining the most-likely-output signal that Semantic Entropy discards. **SAR** (Duan et al., 2024) reweights token log-probabilities by their semantic relevance and adjusts sentence-level uncertainty by cross-sample similarity. **MI** (Yadkori et al., 2024) targets epistemic uncertainty by computing the KL divergence between the empirical joint over responses from iterative prompting and the product of marginals.

3.1.2 Semantic Cluster Features

For ambiguity prediction, we additionally include semantic cluster properties of sampled responses. The cluster features include the number of clusters, cluster assignment entropy, effective number of answers, maximum cluster probability, the probability mass of the top two clusters, and the probability gap between the top two clusters.

We measure the AUROC of each scalar feature as a standalone binary classifier of the ambiguity label, presented in Table 1. For high-dimensional representation features (last-layer embeddings and mid-layer residuals), we report the mean AUROC from 5-fold cross-validated logistic regression after PCA reduction to 128 components.²

Uncertainty features show consistent positive correlation with ambiguity across datasets and models, with entropy-based and cluster-diversity features showing stronger correlations than concentration-based features such as max cluster probability. Individually, however, all scalar UQ and cluster features remain close to chance (≈ 0.5 AUROC) at predicting ambiguity. The representation features are the exception, reaching 0.71–0.73 on AmbigQA. This gap is precisely why the ambiguity classifier (described below) is built on internal representations rather than scalar features.

3.1.3 Internal Representations

Motivated by the body of work discussed in Section 2, we incorporate two representation features, namely the final-layer hidden state embedding and a mid-layer residual (the per-model median layer) from the last generated token, providing complementary views of the model’s internal state at generation time.

²The results for all models are presented in Appendix A, showing the same pattern.

Data	Internal		UQ							Semantic Cluster					
	Emb	Res	SE	SEng	MSP	SAR	MI	CoCoA		#Cl	CIE	MxC	ENA	PG2	Top2
Ambig	.71	.73	.54	.54	.54	.55	.54	.54		.54	.54	.46	.54	.46	.45
NCQA	.49	.60	.50	.52	.57	.55	.54	.53		.59	.50	.52	.50	.52	.43
Trivia	.54	.56	.56	.57	.56	.57	.56	.57		.56	.56	.44	.56	.45	.44

Table 1: AUROC of individual features for predicting ambiguity (Llama-3.1-8B, validation split). Scalar features evaluated directly. Representation features use 5-fold CV logistic regression. Emb = last-token embedding, Res = mid-layer residual. #Cl = number of clusters, CIE = cluster entropy, MxC = max cluster probability, ENA = effective number of answers, PG2 = probability gap (top-2), Top2 = top-2 probability mass.

3.2 The Role of Ambiguity in Error Prediction via UQ

Our framework comprises three modules trained in two stages, given as pseudocode in Algorithm 1. **Stage 1** trains an *Ambiguity Model* that predicts input ambiguity from the LLM’s internal representations. **Stage 2** trains an *error-prediction model* over UQ features whose output is conditioned on the (frozen) Stage-1 ambiguity signal, either by routing (Gated Experts) or by an abstention penalty (Selective Prediction). Gated Experts and Selective Prediction are two *alternative* Stage-2 designs, never combined. Both share the baseline error prediction head (Section 4.4), which maps the UQ feature vector $\mathbf{x}_f$ (the scalar UQ metrics and semantic-cluster statistics of Section 3.1, not the internal representations, which feed only the Ambiguity Model) to a scalar *confidence logit* $\hat{c}$ (higher $\hat{c}$ indicates a more likely correct answer, so the uncertainty score is $-\hat{c}$). The two differ only in how they incorporate the ambiguity signal.

The Ambiguity Model is trained independently of the error-prediction pipeline. Unlike the error-prediction experts, its inputs are the *internal representation features* of Section 3.1, namely the last-layer hidden state $\mathbf{e}$ and the mid-layer residual $\mathbf{r}$ of the final generated token (optionally concatenated with the scalar UQ vector), motivated by Table 1, where representations are by far the strongest standalone predictor of ambiguity. It encodes these into $\mathbf{z}_a = \text{MLP}_a([\mathbf{e}; \mathbf{r}])$ and predicts ambiguity via $\hat{a} = \sigma(\mathbf{w}^\top \mathbf{z}_a + b)$, optimised with binary cross-entropy against gold ambiguity labels (5-run ensemble). The resulting probability $\hat{a}$ is frozen and used as the gating or penalty signal below.

The **Gated Experts** model routes instances to two specialised expert heads, E_u and E_a , each an MLP over the encoded UQ features $\mathbf{x}_f$ that out-

Algorithm 1 Ambiguity-conditioned error prediction

Require: LLM M , UQ metrics, data with correctness labels c and ambiguity labels a

Stage 1: Ambiguity Model over representations

- 1: **for** each instance i **do**
- 2: extract $\mathbf{e}_i$ (last-layer embedding), $\mathbf{r}_i$ (mid-layer residual), $\mathbf{x}_i$ (UQ vector)
- 3: **end for**
- 4: train A : $\hat{a} = \sigma(\mathbf{w}^\top \text{MLP}_a([\mathbf{e}_i; \mathbf{r}_i]) + b)$ with BCE against a . Freeze A
- Stage 2: error prediction over UQ features, conditioned on A**
- 5: compute $\hat{a}_i = A(\mathbf{e}_i, \mathbf{r}_i)$ for all i
- 6: train the error head(s) on $\mathbf{x}_i$ with $\mathcal{L}(\theta) = \mathcal{L}_{\text{BCE}}(\hat{c}, c) + \alpha \mathcal{L}_{\text{rank}}(\hat{c})$:
- 7: **Gated:** $\hat{c}_i = (1-g_i)E_u(\mathbf{x}_i) + g_iE_a(\mathbf{x}_i)$, $g_i = a_i$ (oracle) or $\hat{a}_i$ (latent)
- 8: **Selective:** $\hat{c}_i = \text{head}_c(\mathbf{x}_i)$, ranked by $R_i = -\hat{c}_i + \lambda s_i$, $s_i = a_i$ or $\hat{a}_i$

Ensure: AUROC from confidence $\hat{c}$ and PRR from rejection score R ($-\hat{c}$ for baseline/gated, $-\hat{c} + \lambda s$ for selective)

puts a confidence logit (see bottom of Figure 1). During training each expert is supervised preferentially on the instances routed to it (a specialisation loss weighted by the gate), so E_u captures the error patterns of unambiguous inputs and E_a those of ambiguous inputs, mitigating feature interference. The prediction is the gated mixture $\hat{c} = (1-g)E_u(\mathbf{x}_f) + gE_a(\mathbf{x}_f)$. The **Oracle Version** uses the ground-truth label $g = a \in \{0, 1\}$. The **Latent Version** uses the predicted $g = \hat{a}$.

The **Selective Prediction Framework** (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017) ranks instances for abstention by a rejection score $R = -\hat{c} + \lambda s$ that penalises low-confidence, ambiguous instances (fixed penalty weight $\lambda = 1$, weighting the confidence and ambiguity terms equally rather than tuned). Unlike Cole et al. (2023), who threshold on sampling-based confidence alone, we inject an explicit, supervised ambiguity term, letting the model abstain preferentially on uncertain, ambiguous instances. The **Oracle Version** uses $s = a$ and the **Latent Version** uses $s = \hat{a}$. Where ambiguity is negatively associated with error (NCQA, where either answer to a conflicting-evidence question is accepted), the penalty sign is reversed. Both models use the composite objective $\mathcal{L}(\theta) = \mathcal{L}_{\text{BCE}}(\hat{c}, c) + \alpha \mathcal{L}_{\text{rank}}(\hat{c})$ of Section 4.4. The gated model also applies a per-expert specialisation loss (weight 0.3).

AmbigQA (Train: 10,036 / Dev: 2,002 / Test: -)

Question: The most common type of rock in Earth’s crust is?

Answers: Mafic rocks, Granite

Ambiguity Explanation: Ambiguous referent. Mafic rocks dominate the oceanic crust, while granite is the most common in the continental crust.

TriviaQA (Train: 61,888 / Dev: 7,993 / Test: -)

Question: Marie Curie’s country of birth?

Answers: Poland

Ambiguity Explanation: N/A

NCQA (Train: 234 / Dev: 100 / Test: 100)

Question: Were the Middle Ages the Dark Ages?

Answers: Yes, No

Ambiguity Explanation: Competing evidence. Roman institutional collapse supports the “darkness” narrative, while Islamic and Carolingian peaks demonstrate intellectual progress.

Table 2: Examples and sizes of the datasets. AmbigQA and TriviaQA do not release test sets with answers.

4 Experiments

4.1 Data

The **AmbigQA** dataset (Min et al., 2020) extends NQ (Kwiatkowski et al., 2019) with multiple plausible answer sets. Over half of its natural queries are inherently ambiguous, e.g. through conflicting entity references or temporal dependencies.

The **TriviaQA** dataset (Joshi et al., 2017) is a large-scale reading comprehension benchmark containing question-answer pairs authored by trivia enthusiasts. As discussed in Section 2, while it is not officially an ambiguous question dataset, approximately 16% of the instances in the dataset contain ambiguous questions. We extend the TriviaQA dataset with underspecification annotations following the protocol of Huang et al. (2026), labeling each instance as underspecified or not using a Qwen (Qwen Team, 2025) model with Thinking Mode. Huang et al. (2026) validate the quality of this annotation approach against human judgments.

The **NCQA** dataset (Nachshoni et al., 2025) comprises yes/no questions derived from fact-checking sources. While very small in size, this dataset offers a different source of aleatoric uncertainty, namely conflicting evidence rather than ambiguity in the questions.

4.2 Models

To generate the answers to the questions in the QA datasets, we employ five large language models rep-

representing three distinct paradigms, namely standard instruction tuning, diversity-oriented post-training, and ambiguity-aware fine-tuning.

For our standard baselines, we utilize **Llama-3.1-8B** (Grattafiori et al., 2024) and **Qwen3-14B** (Qwen Team, 2025). We deploy Qwen3-14B in its non-thinking mode to avoid test-time compute scaling.

To investigate interventions for aleatoric uncertainty, we evaluate **Spectrum-Llama-3.1-8B-v1** and **Spectrum-Qwen3-14B-v1** (Sorensen et al., 2025). These variants undergo Spectrum tuning to mitigate mode collapse, preserving distributional coverage and generating diverse, valid responses to ambiguous prompts.

Finally, we benchmark against **A²Search-7B-Instruct** (Zhang et al., 2025a), a model explicitly fine-tuned via reinforcement learning to recognize ambiguity, navigate conflicting evidence, and directly resolve underspecified queries. This model is based on Qwen2.5-7B (Qwen Team, 2024).

4.3 Output Evaluation

To evaluate model performance, we employ an LLM-as-a-judge paradigm (Zheng et al., 2023), using Gemma-2-2B-IT (Gemma Team, 2024). We choose Gemma to keep the judge architecturally independent from the Llama- and Qwen-based models being evaluated, avoiding same-family bias, while remaining small enough to score the full evaluation set efficiently. The judge is prompted to determine if the generated prediction is factually equivalent to the gold reference answers, allowing for minor surface-level variations. This verification is treated as a binary classification task, marking the prediction as correct if the normalized softmax probability of the generated “yes” token exceeds a predefined threshold ($\tau = 0.5$).

To evaluate factual consistency, we score responses using AlignScore (Zha et al., 2023), which has been trained to evaluate factual correctness of a predicted answer against the gold. For instances containing multiple valid reference answers, we follow the common practice of evaluating the prediction against each reference independently, and assigning the maximum AlignScore achieved across all references for that instance (Min et al., 2020; Joshi et al., 2017).

4.4 Error Prediction Baseline

The base model takes as input either a single scalar UQ metric or a combination of scalar UQ met-

rics $\mathbf{x}_f$, processed through a multi-layer perceptron (MLP) with layer normalisation and dropout to obtain a latent representation $\mathbf{z} = \text{MLP}(\mathbf{x}_f)$. This representation is then passed to the **Error Prediction** head, which estimates factual error likelihood via a non-linear classifier $\hat{c} = \text{head}_c(\mathbf{z})$. The model parameters θ , comprising the weights of the MLP and the error head, are optimized with a composite objective $\mathcal{L}(\theta) = \mathcal{L}_{\text{BCE}}(\hat{c}, c) + \alpha \mathcal{L}_{\text{rank}}(\hat{c})$, where the predicted confidence $\hat{c} = \text{head}_c(\mathbf{z})$ depends on θ , $\mathcal{L}_{\text{BCE}}(\hat{c}, c)$ is the binary cross-entropy (BCE) loss and $\mathcal{L}_{\text{rank}}(\hat{c}) = \frac{1}{|P||N|} \sum_{i \in P, j \in N} \log(1 + \exp(\hat{c}_j - \hat{c}_i))$ is a pairwise softplus ranking loss over correct (P) and incorrect (N) samples with $\alpha = 0.05$, ensuring monotonic confidence ordering.

4.5 Error Prediction Evaluation

To evaluate the discriminative power of our uncertainty estimates, we utilize the **Area Under the Receiver Operating Characteristic curve (AUROC)** (Fawcett, 2006). AUROC provides a threshold-independent measure of a model’s ability to rank incorrect generations as more uncertain than correct ones. Additionally, we report the **Prediction Rejection Ratio (PRR)** (Malinin and Gales, 2020), which measures how effectively uncertainty scores drive abstention. It normalizes the area under the accuracy-rejection curve against a random baseline and an optimal oracle. For all variants the uncertainty (rejection) score is $R = -\hat{c}$ for the baseline and gated models and $R = -\hat{c} + \lambda s$ for the selective model. AUROC is computed on the confidence $\hat{c}$ and PRR on R . Statistical significance of improvements over the baseline is assessed using a paired bootstrap over instances ($p < 0.05$).

4.6 Implementation Details

To quantify semantic uncertainty, we process each input in two generation phases. First, we establish a baseline response via greedy decoding, extracting both its sequence log-likelihood and the mid-layer residual of its final generated token. Next, we draw $K = 10$ stochastic samples $\{a_k\}_{k=1}^K$ using multinomial sampling ($T = 1.0$).

The experiments were carried out on NVIDIA GH200 Grace Hopper Superchips. Error Prediction models were trained for a maximum of 100 epochs, with early stopping when PRR on the tuning subset (described below) does not improve for 10 epochs. We carried out a hyperparameter search with hidden dimensions in $\{128, 256, 512, 1024, 2048\}$

and learning rates in $\{3 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$. To keep model selection strictly separate from evaluation, we never tune on the set we report. For AmbigQA and TriviaQA, whose official test labels are withheld, we carve a 15% tuning subset out of the training split, stratified by ambiguity and correctness with a fixed seed. Every selection decision, namely the hidden dimension, the learning rate and the early-stopping epoch, is made on this tuning subset, and the official validation split is then used exactly once, for final reporting. NCQA has a genuine held-out test split, which we report on while selecting on its training data. This guarantees that no tuning decision is ever informed by the split on which a number is reported.

5 Results

5.1 QA Performance

Table 3 presents the LLM-as-a-judge and AlignScore values across models on the validation sets of all three datasets. All models score higher on TriviaQA and NCQA datasets as they are easier, given that TriviaQA is mostly unambiguous, and NCQA is a binary yes/no question set, while AmbigQA proves to be a much harder dataset to correctly classify. Comparing ambiguous and unambiguous instances, on AmbigQA and TriviaQA all models score higher on unambiguous instances, indicating these are genuinely harder despite the evaluation metrics taking the maximum score over all plausible reference answers, which if anything advantages the ambiguous ones. NCQA is the exception (Table 3), where its ambiguous conflicting-evidence questions accept either answer, so models score at ceiling on them. AlignScore values are lower across the board, indicating the stricter judgements of this metric compared to LLM-as-a-judge. For instance, while LLM-as-a-judge accepts the prediction ‘2001’ as a correct approximation of the reference ‘2001 fiscal year’, AlignScore deems it not sufficiently precise.

5.2 Ambiguity Prediction

Table 4 presents the ambiguity prediction scores for different models and datasets. Unsurprisingly, ambiguity prediction is most successful on AmbigQA, which has high quality human annotated ambiguity labels. In contrast, TriviaQA relies on LLM-generated annotations for underspecification (as discussed in Section 4.1). Furthermore, the limited scale of NCQA restricts the models’ ability to

Model	Dataset	All		Unamb.		Amb.	
		LLM	Align	LLM	Align	LLM	Align
Qwen3 14B	AmbigQA	0.526	0.332	0.548	0.346	0.497	0.315
	TriviaQA	0.719	0.706	0.759	0.744	0.675	0.666
	NCQA	0.920	0.920	0.556	0.556	1.000	1.000
Llama 3.1-8B	AmbigQA	0.618	0.434	0.631	0.457	0.602	0.405
	TriviaQA	0.829	0.810	0.857	0.838	0.799	0.781
	NCQA	0.850	0.850	0.167	0.167	1.000	1.000
Spectrum-Llama 3.1-8B	AmbigQA	0.662	0.403	0.673	0.419	0.649	0.382
	TriviaQA	0.809	0.783	0.838	0.813	0.779	0.751
	NCQA	0.860	0.860	0.222	0.222	1.000	1.000
Spectrum-Qwen3 14B	AmbigQA	0.574	0.411	0.588	0.427	0.557	0.391
	TriviaQA	0.769	0.751	0.806	0.786	0.729	0.713
	NCQA	0.910	0.910	0.500	0.500	1.000	1.000
A ² Search 7B	AmbigQA	0.466	0.281	0.477	0.287	0.452	0.274
	TriviaQA	0.625	0.624	0.670	0.663	0.577	0.583
	NCQA	0.910	0.910	0.500	0.500	1.000	1.000

Table 3: LLM-judge (LLM) and AlignScore (Align) accuracy on validation splits by ambiguity (test for NCQA).

Model	AmbigQA	TriviaQA	NCQA
Llama-3.1-8B	0.723	0.621	0.626
Qwen3-14B	0.728	0.622	0.618
Spectrum-Llama-3.1-8B	0.730	0.622	0.647
Spectrum-Qwen3-14B	0.732	0.624	0.614
A ² Search-7B	0.693	0.623	0.642

Table 4: Ambiguity prediction AUROC per model and dataset (validation split, test for NCQA).

extract robust patterns.

5.3 Error Prediction

The difference in UQ metric ability to predict model errors is presented in Table 6.³ We find that the ambiguity signal improves error prediction, with the gains concentrated on the UQ metrics that most entangle aleatoric and epistemic uncertainty (Semantic Entropy and Semantic Energy). Gold ambiguity (*oracle*) helps where the labels are reliable (AmbigQA, NCQA), while predicted ambiguity (*latent*) helps where the learned signal carries information beyond the labels. On TriviaQA it even surpasses its oracle counterpart, whereas on AmbigQA it adds less to the already-strong *all UQ* representation.

Figure 3 illustrates the improvements from our proposed models. Each curve traces the error rate among accepted answers as we accept progressively more of them, ordered from most to least confident, so a well-calibrated measure keeps errors low until coverage is high. Curves are shown separately for ambiguous and unambiguous questions, with upper-bound lines for a measure that ranked every incorrect answer last. First, the improvement

³The results on all other models are in Appendix B.

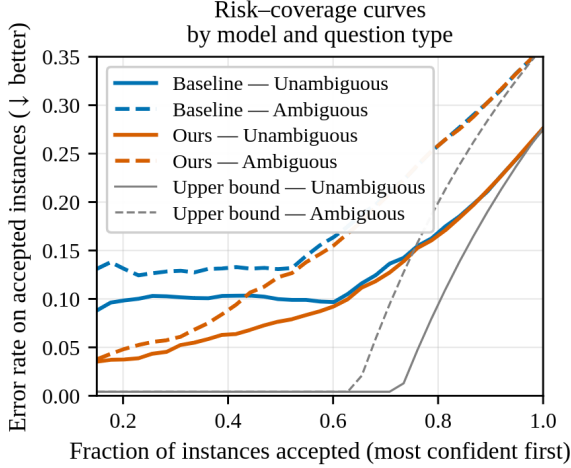


Figure 3: Risk-coverage curves for baseline and latent selective models on unambiguous and ambiguous questions, with upper bound (Qwen3-14B, TriviaQA).

appears on both ambiguous and unambiguous instances, showing that the architecture encodes both relationships, between epistemic uncertainty and error likelihood (unambiguous) and between combined aleatoric and epistemic uncertainty and error likelihood (ambiguous). The full per-subset results across all models and datasets are provided in Appendix B, confirming this pattern holds beyond the model shown here. Second, the improvement is most pronounced in the high-confidence area, where mistakes are the most detrimental.

Nonetheless, interesting differences emerge between the UQ metrics themselves. Among the individual UQ features, CoCoA is the strongest single metric. The largest gains from ambiguity, however, appear on the weaker, more entangled metrics (Semantic Entropy and Semantic Energy) rather than on the already-strong single metrics, consistent with ambiguity being most useful where the base metric fails to separate aleatoric from epistemic uncertainty. Conversely, on the hardest datasets (AmbigQA and NCQA) the routing and the selective penalty ($\lambda = 1$) can overwhelm the *weakest* base metrics, SAR and MSP, degrading their PRR even with gold ambiguity labels. The benefit of ambiguity is therefore concentrated on metrics that carry enough correctness signal to condition on, rather than on those already close to chance. Contrary to the design of MI as a UQ metric which is aware of aleatoric ambiguity, we find that the metric performs very poorly, especially on AmbigQA, corroborating the findings of previous research (Tomov et al., 2025). Error prediction is

Configuration	AmbigQA		TriviaQA	
	all UQ	Sem. Ent.	all UQ	Sem. Ent.
Baseline (no ambiguity)	0.464	0.444	0.785	0.641
Full method	0.474	0.496	0.811	0.762
– ranking loss	0.471	0.484	0.811	0.758
– specialisation loss	0.482	0.476	0.809	0.754
– representation feats	0.485	0.485	0.810	0.758
– ambiguity penalty	0.491	0.477	0.809	0.757
binary conditioning	0.468	0.452	0.800	0.713

Table 5: Controlled component ablation on Qwen3-14B (latent-selective PRR, AlignScore, held-out; hidden size 1024, learning rate 10^{-4}). Each row removes one design choice while holding the others at the full setting. On *all UQ* the components stay within ± 0.02 seed noise; on Semantic Entropy every component helps and *binarising* the ambiguity signal (last row) is by far the most damaging.

strongest when combining all UQ features (*all UQ*). Even then the ambiguity signal yields further improvements where it is not redundant with it (latent on TriviaQA and NCQA, gold on AmbigQA and NCQA), supporting our hypothesis that UQ metrics, even combined, do not fully separate aleatoric from epistemic uncertainty.

A controlled ablation isolates each design choice on Qwen3-14B (Table 5). The effect is only measurable where the ambiguity mechanism has room to help: on the strong combined feature (*all UQ*) removing any single component stays within single-seed noise (± 0.02), whereas on Semantic Entropy every component contributes. The dominant factor is *continuous* conditioning: binarising the predicted ambiguity into a hard ambiguous/unambiguous split costs about 0.05 PRR on both datasets (about 40% of the gain), far more than the ranking loss, specialisation loss, representation features or ambiguity penalty, which each contribute between 0.005 and 0.02. The benefit thus comes from conditioning on a *graded* ambiguity signal rather than a hard split.

The NCQA improvements are the most dramatic because its ambiguous cases are the easy ones, as both answers to a binary question are accepted. These gains should be read with caution given NCQA’s small training set of 234 instances, but they are on the held-out test split and consistent across feature sets and models, so they are unlikely to be an overfitting artefact.

Finally, models finetuned to represent multiple plausible answers (e.g. Spectrum-Qwen) barely outperform their vanilla counterparts at error prediction, suggesting such finetuning improves calibration.

Dataset	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
AmbigQA	all UQ	0.483	0.491	0.535	0.459	0.477
	CoCoA	0.538	0.534	0.550	0.467	0.527
	Sem. energy	0.393	0.435	0.464	0.427	0.433
	Sem. entropy	0.445	0.481	0.487	0.448	0.456
	SAR	0.532	0.254	0.101	0.147	0.122
	MSP	0.146	0.010	0.157	-0.041	-0.117
	MI	0.014	0.038	0.058	0.038	0.051
TriviaQA	all UQ	0.786	0.795	0.785	0.811	0.774
	CoCoA	0.753	0.769	0.753	0.782	0.741
	Sem. energy	0.754	0.775	0.756	0.792	0.744
	Sem. entropy	0.641	0.759	0.672	0.762	0.669
	SAR	0.735	0.771	0.738	0.784	0.726
	MSP	0.738	0.758	0.737	0.770	0.726
	MI	0.203	0.572	0.177	0.572	0.182
NCQA	all UQ	0.499	0.577	0.967	0.534	0.973
	CoCoA	0.476	0.606	0.963	0.520	0.964
	Sem. energy	0.468	0.489	0.971	0.436	0.972
	Sem. entropy	0.086	0.533	0.968	-0.545	0.956
	SAR	0.684	0.644	0.984	-0.319	0.986
	MSP	0.444	0.623	0.951	0.645	0.971
	MI	-0.160	0.576	0.946	0.576	0.945

Table 6: PRR on validation split (test for NCQA) for Qwen3-14B (AlignScore supervision). **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best PRR for model/dataset.

tion alongside steerability but leaves room for further gains. On these models the explicit ambiguity signal adds little (latent gating is within noise of the baseline), consistent with the finetuning already encoding part of the ambiguity structure that the signal would otherwise supply.

To illustrate, consider the pair of questions in Table 7. The top question is underspecified, where valid answers include 256 (total levels) and 21 (levels before difficulty plateaus). The question about sonnets has one correct answer (14 lines). A UQ-based model trained on semantic entropy treats output diversity as a signal of likely error, assigning a high error score (73.8%) to the ambiguous question. Our latent selective model disentangles underspecification from error, assigning a near-zero rejection score (0.0%) while giving a similarly low score for the unambiguous sonnet (1.2%). The ambiguity model makes correct predictions (0.527 for Pac-Man versus 0.440 for the sonnet), though the small margin reflects the difficulty of predicting ambiguity discussed in Section 5.2.

6 Conclusion

In this paper we claim that ambiguity signal is useful for predicting errors from uncertainty, because uncertainty encompasses both epistemic and aleatoric sources. Experimental results show that ambiguity information is beneficial across models, feature sets and architecture choices for error

Amb.	Question	Acc	Base	Ours	$\hat{p}_{\text{amb}}$
✓	How many levels are there in Pac-Man?	1.00	0.74	0.00	0.53
✗	How many lines are there in a sonnet?	1.00	0.03	0.01	0.44

Table 7: Uncertainty scores used for ranking (normalised). Base = $-\hat{c}$ from semantic entropy. Ours = the latent-selective rejection score $R = -\hat{c} + \lambda\hat{a}$, which differs from Base only through the ambiguity penalty (the two share a confidence head). Acc = AlignScore of the most likely response, and $\hat{p}_{\text{amb}}$ = predicted ambiguity.

prediction, with the gains concentrated on the UQ metrics that most entangle aleatoric and epistemic uncertainty. We show that using UQ metrics along with ambiguity can yield high error prediction performance, up to 0.880 PRR and 0.907 AUROC on TriviaQA. These results contribute to improving the reliability of LLMs.

Limitations

The study covers only textual data, and all questions and answers are in English, so further work is needed to establish whether the trends generalise to other languages, multilingual models and multimodal models.

Moreover, our latent models require ambiguity-labelled data at training time, which may not be available for all domains and tasks. Inducing ambiguity labels fully automatically remains an important direction for future work.

Furthermore, we focus on semantic uncertainty quantification metrics. These achieve state-of-the-art results and are therefore the most interesting to study, but whether other UQ metrics share the same relationship to ambiguity remains open.

Finally, we study only closed-book QA, where the model relies solely on parametric knowledge. Extending the work to retrieval-augmented generation, especially cases where retrieved evidence contradicts internal knowledge, is an interesting direction.

Acknowledgments

Ieva Raminta Staliūnaitė is supported by a research grant from Imminent, Translated Research Center, and the Alan Turing Institute Enrichment Scheme. Andreas Vlachos is supported by the DARPA program SciFy.

References

- Gustaf Ahndritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. 2024. [Distinguishing the knowable from the unknowable with language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. [Interpreting predictive probabilities: Model confidence or human label variation?](#) In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian’s, Malta. Association for Computational Linguistics.
- Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. [Do androids know they’re only dreaming of electric sheep?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4401–4420, Bangkok, Thailand. Association for Computational Linguistics.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. [Inside: LLMs’ internal states retain the power of hallucination detection](#). *ArXiv*, abs/2402.03744.
- Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. [Selectively answering ambiguous questions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.
- Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. 2025. [Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23696–23703.
- Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. [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)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.
- Ran El-Yaniv and Yair Wiener. 2010. [On the foundations of noise-free selective classification](#). *Journal of Machine Learning Research*, 11(53):1605–1641.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630(8017):625–630.
- Tom Fawcett. 2006. [An introduction to roc analysis](#). *Pattern Recognition Letters*, 27(8):861–874. ROC Analysis in Pattern Recognition.
- Yonatan Geifman and Ran El-Yaniv. 2017. [Selective classification for deep neural networks](#). *Advances in neural information processing systems*, 30.
- Gemma Team. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Dan Hendrycks and Kevin Gimpel. 2016. [A baseline for detecting misclassified and out-of-distribution examples in neural networks](#). *arXiv preprint arXiv:1610.02136*.
- Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. [Decomposing uncertainty for large language models through input clarification ensembling](#). In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Yunchong Huang, Gianni Barlacchi, and Sandro Pezzelle. 2026. [Who is the richest club in the championship? detecting and rewriting underspecified questions improve qa performance](#). *arXiv preprint arXiv:2602.11938*.
- Eyke Hüllermeier and Willem Waegeman. 2021. [Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods](#). *Machine learning*, 110(3):457–506.
- Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025. [Exploring concept depth: How large language models acquire knowledge and concept at different layers?](#) In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 558–573, Abu Dhabi, UAE. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [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)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. [Language models \(mostly\) know what they know](#). *arXiv preprint arXiv:2207.05221*.
- Alex Kendall and Yarin Gal. 2017. [What uncertainties do we need in bayesian deep learning for computer vision?](#) *Advances in Neural Information Processing Systems*, 30.

- 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. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. [Generating with confidence: Uncertainty quantification for black-box large language models](#). *arXiv preprint arXiv:2305.19187*.
- Zhu Liu and Ying Liu. 2023. [Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3963–3977, Toronto, Canada. Association for Computational Linguistics.
- Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. 2025. [Semantic energy: Detecting llm hallucination beyond entropy](#). *arXiv preprint arXiv:2508.14496*.
- Andrey Malinin and Mark Gales. 2020. [Uncertainty estimation in autoregressive structured prediction](#). *arXiv preprint arXiv:2002.07650*.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Eviatar Nachshoni, Arie Cattán, Shmuel Amar, Ori Shapira, and Ido Dagan. 2025. [Consensus or conflict? fine-grained evaluation of conflicting answers in question-answering](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 138–159, Suzhou, China. Association for Computational Linguistics.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. [Llms know more than they show: On the intrinsic representation of llm hallucinations](#). *ArXiv*, abs/2410.02707.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. [Trust me, I’m wrong: LLMs hallucinate with certainty despite knowing the answer](#). *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14665–14688.
- Taylor Sorensen, Benjamin Newman, Jared Moore, Chan Park, Jillian Fisher, Niloofar Mireshghal, Liwei Jiang, and Yejin Choi. 2025. [Spectrum tuning: Post-training for distributional coverage and in-context steerability](#). *arXiv preprint arXiv:2510.06084*.
- Ieva Raminta Staliūnaitė, Julius Cheng, and Andreas Vlachos. 2026. [Uncertainty quantification for evaluating gender bias in machine translation](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2204–2225, Rabat, Morocco. Association for Computational Linguistics.
- Ieva Raminta Staliūnaitė and Andreas Vlachos. 2025. [Uncertain \(mis\)takes at LeWiDi-2025: Modeling human label variation with semantic entropy](#). In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 256–262, Suzhou, China. Association for Computational Linguistics.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. 2025. [The illusion of certainty: Uncertainty quantification for llms fails under ambiguity](#). *arXiv preprint arXiv:2511.04418*.

- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. [Cocoa: A minimum bayes risk framework bridging confidence and consistency for uncertainty quantification in llms](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 106236–106281. Curran Associates, Inc.
- Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. [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)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.
- Nassim Walha, Sebastian G Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. 2026. [Fine-grained uncertainty decomposition in large language models: A spectral approach](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26090–26098.
- Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. [To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty](#). *Advances in Neural Information Processing Systems*, 37:58077–58117.
- Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. [MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5861–5878, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating factual consistency with a unified alignment function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.
- Fengji Zhang, Xinyao Niu, Chengyang Ying, Guancheng Lin, Zhongkai Hao, Zhou Fan, Chengen Huang, Jacky Keung, Bei Chen, and Junyang Lin. 2025a. [A² search: Ambiguity-aware question answering with reinforcement learning](#). *arXiv preprint arXiv:2510.07958*.
- Zhuoxuan Zhang, Jinhao Duan, Edward Kim, and Kaidi Xu. 2025b. [Sparse neurons carry strong signals of question ambiguity in LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16081–16099, Suzhou, China. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Full Feature Correlation with Ambiguity Labels

Table 8 extends the per-feature AUROC analysis of Table 1 to the remaining four models. The pattern observed for Llama-3.1-8B holds across model families. Representation features (last-layer embeddings and mid-layer residuals) provide the strongest standalone signal for ambiguity on AmbigQA, while scalar UQ and semantic-cluster features cluster near chance, with entropy- and diversity-based features slightly outperforming concentration-based ones.

Model	Dataset	Internal Repr.		UQ					Sem. Clust.						
		Emb	Res	SE	SEng	MSP	SAR	MI	CoCoA	#CI	CIEnt	MxCP	ENA	PG2	Top2
Spectrum-Llama 3.1-8B	AmbigQA	0.726	0.741	0.541	0.538	0.529	0.510	0.542	0.532	0.536	0.541	0.459	0.541	0.461	0.460
	NCQA	0.501	0.603	0.507	0.490	0.572	0.519	0.541	0.508	0.556	0.507	0.491	0.507	0.494	0.465
	TriviaQA	0.576	0.570	0.560	0.566	0.556	0.551	0.565	0.563	0.565	0.561	0.443	0.561	0.448	0.437
Qwen3 14B	AmbigQA	0.718	0.715	0.525	0.507	0.484	0.477	0.458	0.522	0.518	0.527	0.469	0.527	0.468	0.480
	NCQA	0.432	0.442	0.450	0.495	0.503	0.464	0.514	0.499	0.481	0.478	0.522	0.478	0.522	0.500
	TriviaQA	0.588	0.594	0.546	0.555	0.554	0.555	0.463	0.556	0.549	0.548	0.453	0.548	0.455	0.454
Spectrum-Qwen3 14B	AmbigQA	0.730	0.733	0.538	0.524	0.519	0.501	0.541	0.526	0.535	0.538	0.463	0.538	0.464	0.462
	NCQA	0.305	0.401	0.463	0.457	0.528	0.474	0.505	0.513	0.568	0.463	0.566	0.463	0.577	0.402
	TriviaQA	0.571	0.579	0.565	0.567	0.562	0.539	0.567	0.570	0.566	0.565	0.439	0.565	0.443	0.436
A ² Search 7B	AmbigQA	0.707	0.714	0.494	0.476	0.451	0.458	0.460	0.483	0.485	0.494	0.502	0.494	0.500	0.507
	NCQA	0.634	0.514	0.515	0.546	0.608	0.552	0.491	0.598	0.485	0.486	0.514	0.486	0.514	0.513
	TriviaQA	0.588	0.584	0.540	0.545	0.542	0.543	0.468	0.543	0.540	0.539	0.462	0.539	0.464	0.468

Table 8: Same as Table 1 for Spectrum-Llama, Qwen3-14B, Spectrum-Qwen3-14B and A²Search-7B. AUROC of individual features for predicting ambiguity across all models and datasets (validation split). Scalar features evaluated directly. Representation features use 5-fold CV logistic regression.

B Full Error Prediction Results

Tables 9–20 report PRR and AUROC for error prediction across all five models, three datasets, and two correctness supervision signals (AlignScore and LLM-judge), with per-subset scores for ambiguous and unambiguous instances shown beneath each main value. Across this full grid, the pattern reported in Section 5.3 for Qwen3-14B holds. Ambiguity-aware models improve over the UQ-only baseline for most model and feature combinations, most consistently on the entangled metrics (Semantic Entropy and Semantic Energy) and on the combined representation. On TriviaQA the latent variants track or exceed the oracle. On AmbigQA and NCQA the oracle is stronger. Degradations concentrate on the weakest single metrics (SAR and MSP) on AmbigQA and NCQA, where the routing and the selective penalty overwhelm a base signal that is already close to chance, and on MI.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.575	0.568	0.609	0.556	0.582
		0.657	0.667	0.674	0.635	0.657
		0.444	0.441	0.412	0.424	0.438
	CoCoA	0.615	0.607	0.627	0.579	0.606
		0.705	0.693	0.706	0.657	0.706
		0.490	0.470	0.490	0.440	0.490
	Semantic energy	0.593	0.587	0.606	0.555	0.582
		0.683	0.669	0.682	0.633	0.682
		0.462	0.453	0.461	0.422	0.461
	Semantic entropy	0.592	0.590	0.610	0.580	0.584
		0.680	0.673	0.680	0.643	0.680
		0.463	0.462	0.462	0.440	0.462
	SAR	0.514	0.308	0.441	0.145	0.458
		-0.285	-0.346	0.481	-0.280	0.527
		-0.177	-0.139	0.459	-0.058	0.360
	MSP	0.390	0.316	0.422	0.196	0.325
		-0.566	0.207	0.277	0.176	0.348
		-0.367	0.094	-0.039	0.064	0.244
	MI	0.176	0.151	0.187	0.146	0.182
		0.156	0.146	0.139	0.130	0.158
		0.127	0.134	0.118	0.130	0.123
Qwen3-14B	all UQ	0.483	0.491	0.535	0.459	0.477
		0.556	0.594	0.596	0.556	0.572
		0.366	0.332	0.376	0.291	0.369
	CoCoA	0.538	0.534	0.550	0.467	0.527
		0.651	0.627	0.650	0.577	0.650
		0.377	0.325	0.379	0.288	0.379
	Semantic energy	0.393	0.435	0.464	0.427	0.433
		0.506	0.548	0.527	0.540	0.505
		0.255	0.227	0.299	0.217	0.255
	Semantic entropy	0.445	0.481	0.487	0.448	0.456
		0.538	0.587	0.538	0.550	0.538
		0.321	0.310	0.321	0.283	0.308
	SAR	0.532	0.254	0.101	0.147	0.122
		-0.429	0.238	0.073	0.056	0.136
		-0.259	0.054	-0.187	-0.041	0.077
	MSP	0.146	0.010	0.157	-0.041	-0.117
		-0.279	-0.192	-0.386	-0.086	-0.285
		-0.145	-0.052	0.162	-0.018	-0.149
	MI	0.014	0.038	0.058	0.038	0.051
		-0.054	0.034	-0.043	0.034	-0.054
		0.015	0.005	0.015	0.006	0.015
Spectrum-Llama-3.1-8B	all UQ	0.460	0.455	0.470	0.449	0.450
		0.507	0.513	0.508	0.481	0.503
		0.386	0.380	0.378	0.355	0.388
	CoCoA	0.465	0.458	0.465	0.432	0.427
		0.503	0.496	0.505	0.469	0.504
		0.403	0.389	0.403	0.371	0.403
	Semantic energy	0.455	0.452	0.459	0.393	0.407
		0.507	0.507	0.509	0.439	0.509
		0.368	0.366	0.370	0.320	0.370
	Semantic entropy	0.455	0.447	0.461	0.420	0.381
		0.515	0.493	0.515	0.470	0.515
		0.371	0.363	0.371	0.340	0.371
	SAR	0.461	0.454	0.464	0.440	0.443
		0.508	0.500	0.508	0.485	0.508
		0.386	0.375	0.386	0.360	0.386
	MSP	0.417	0.413	0.417	0.371	0.379
		0.451	0.441	0.450	0.401	0.450
		0.363	0.355	0.365	0.314	0.364
	MI	0.232	0.191	0.213	0.193	0.212
		0.188	0.200	0.160	0.195	0.180
		0.140	0.181	0.128	0.188	0.134
Spectrum-Qwen3-14B	all UQ	0.559	0.560	0.568	0.553	0.543
		0.599	0.601	0.609	0.597	0.599
		0.504	0.499	0.503	0.493	0.501
	CoCoA	0.537	0.535	0.535	0.514	0.491
		0.567	0.561	0.567	0.546	0.567
		0.492	0.476	0.493	0.456	0.493
	Semantic energy	0.507	0.513	0.519	0.490	0.490
		0.557	0.562	0.558	0.531	0.558
		0.428	0.432	0.431	0.415	0.431
	Semantic entropy	0.544	0.539	0.548	0.518	0.507
		0.580	0.578	0.583	0.558	0.581
		0.480	0.474	0.479	0.456	0.479
	SAR	0.510	0.510	0.516	0.500	0.491
		0.565	0.564	0.565	0.556	0.565
		0.429	0.417	0.429	0.406	0.429
	MSP	0.493	0.491	0.489	0.459	0.453
		0.519	0.514	0.516	0.492	0.516
		0.454	0.440	0.453	0.408	0.453
	MI	0.258	0.213	0.248	0.209	0.242
		0.189	0.205	0.189	0.193	0.198
		0.155	0.202	0.145	0.185	0.146
A2Search-7B	all UQ	0.514	0.505	0.554	0.481	0.526
		0.627	0.607	0.627	0.584	0.614
		0.366	0.354	0.373	0.331	0.355
	CoCoA	0.497	0.511	0.536	0.505	0.513
		0.606	0.596	0.606	0.587	0.606
		0.368	0.381	0.368	0.375	0.368
	Semantic energy	0.405	0.446	0.490	0.446	0.465
		0.527	0.551	0.544	0.547	0.527
		0.290	0.298	0.361	0.302	0.291
	Semantic entropy	0.469	0.505	0.524	0.494	0.490
		0.589	0.598	0.589	0.588	0.589
		0.340	0.360	0.339	0.348	0.340
	SAR	0.487	0.396	0.388	0.287	0.382
		0.472	0.435	0.377	-0.130	0.488
		0.299	0.270	0.331	-0.114	0.309
	MSP	0.200	-0.078	0.149	0.030	-0.116
		-0.389	-0.165	-0.398	-0.088	-0.261
		-0.268	-0.103	0.111	-0.051	-0.107
	MI	0.025	0.087	0.056	0.084	0.048
		-0.003	0.090	0.022	0.087	0.022
		0.040	0.009	-0.003	0.006	0.005

Table 9: PRR on AmbigQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.848	0.849	0.848	0.850	0.839
		0.858	0.859	0.857	0.859	0.858
		0.832	0.834	0.833	0.835	0.832
	CoCoA	0.829	0.830	0.830	0.832	0.817
		0.838	0.838	0.838	0.840	0.838
		0.814	0.814	0.814	0.818	0.814
	Semantic energy	0.814	0.816	0.815	0.819	0.803
		0.822	0.823	0.822	0.824	0.822
		0.799	0.802	0.799	0.804	0.799
	Semantic entropy	0.813	0.820	0.812	0.821	0.801
		0.815	0.826	0.815	0.826	0.815
		0.800	0.806	0.800	0.807	0.801
	SAR	0.824	0.825	0.824	0.825	0.810
		0.832	0.832	0.832	0.830	0.832
		0.808	0.810	0.808	0.812	0.808
	MSP	0.793	0.796	0.793	0.802	0.782
		0.804	0.808	0.804	0.813	0.804
		0.774	0.778	0.774	0.784	0.774
	MI	0.745	0.756	0.752	0.762	0.742
		0.750	0.743	0.751	0.748	0.751
		0.729	0.758	0.728	0.762	0.729
Qwen3-14B	all UQ	0.786	0.795	0.785	0.811	0.774
		0.791	0.797	0.790	0.811	0.790
		0.774	0.785	0.774	0.804	0.773
	CoCoA	0.753	0.769	0.753	0.782	0.741
		0.757	0.774	0.757	0.786	0.757
		0.742	0.757	0.741	0.769	0.742
	Semantic energy	0.754	0.775	0.756	0.792	0.744
		0.764	0.778	0.764	0.792	0.759
		0.737	0.765	0.737	0.784	0.734
	Semantic entropy	0.641	0.759	0.672	0.762	0.669
		0.646	0.757	0.644	0.760	0.644
		0.628	0.751	0.637	0.753	0.637
	SAR	0.735	0.771	0.738	0.784	0.726
		0.736	0.769	0.737	0.781	0.736
		0.726	0.763	0.726	0.777	0.724
	MSP	0.738	0.758	0.737	0.770	0.726
		0.741	0.763	0.741	0.775	0.741
		0.728	0.745	0.728	0.755	0.728
	MI	0.203	0.572	0.177	0.572	0.182
		-0.024	0.565	-0.024	0.565	-0.014
		0.004	0.555	0.003	0.555	0.004
Spectrum-Llama-3.1-8B	all UQ	0.830	0.831	0.831	0.835	0.822
		0.845	0.846	0.845	0.850	0.844
		0.808	0.809	0.809	0.812	0.810
	CoCoA	0.804	0.807	0.805	0.810	0.794
		0.817	0.819	0.817	0.821	0.817
		0.784	0.788	0.784	0.791	0.784
	Semantic energy	0.800	0.806	0.801	0.813	0.792
		0.809	0.815	0.809	0.823	0.809
		0.784	0.790	0.784	0.795	0.784
	Semantic entropy	0.799	0.810	0.803	0.815	0.793
		0.810	0.819	0.810	0.825	0.810
		0.779	0.795	0.779	0.798	0.779
	SAR	0.804	0.808	0.805	0.812	0.796
		0.818	0.822	0.818	0.825	0.818
		0.784	0.786	0.784	0.792	0.784
	MSP	0.745	0.755	0.747	0.768	0.738
		0.761	0.769	0.761	0.785	0.761
		0.722	0.730	0.722	0.742	0.722
	MI	0.722	0.747	0.710	0.754	0.696
		0.719	0.742	0.720	0.753	0.714
		0.707	0.739	0.662	0.743	0.711
Spectrum-Qwen3-14B	all UQ	0.857	0.857	0.859	0.861	0.849
		0.856	0.859	0.863	0.865	0.858
		0.849	0.848	0.849	0.850	0.849
	CoCoA	0.836	0.839	0.837	0.840	0.825
		0.841	0.843	0.841	0.843	0.841
		0.825	0.827	0.825	0.829	0.825
	Semantic energy	0.828	0.831	0.828	0.834	0.821
		0.834	0.837	0.834	0.838	0.834
		0.815	0.819	0.815	0.822	0.815
	Semantic entropy	0.826	0.830	0.831	0.834	0.822
		0.832	0.832	0.833	0.837	0.833
		0.821	0.821	0.820	0.824	0.821
	SAR	0.840	0.842	0.841	0.843	0.831
		0.841	0.843	0.841	0.844	0.841
		0.833	0.835	0.833	0.834	0.833
	MSP	0.788	0.794	0.790	0.804	0.779
		0.794	0.801	0.794	0.812	0.794
		0.774	0.780	0.774	0.786	0.774
	MI	0.765	0.785	0.747	0.786	0.741
		0.728	0.775	0.720	0.776	0.736
		0.752	0.787	0.745	0.785	0.750
A2Search-7B	all UQ	0.741	0.754	0.745	0.772	0.735
		0.767	0.778	0.767	0.795	0.765
		0.708	0.722	0.708	0.738	0.707
	CoCoA	0.692	0.732	0.697	0.741	0.685
		0.720	0.757	0.720	0.767	0.720
		0.656	0.694	0.656	0.701	0.656
	Semantic energy	0.727	0.750	0.734	0.764	0.727
		0.760	0.777	0.759	0.792	0.760
		0.687	0.712	0.688	0.723	0.687
	Semantic entropy	0.658	0.742	0.667	0.745	0.661
		0.678	0.767	0.678	0.770	0.678
		0.629	0.705	0.625	0.707	0.629
	SAR	0.704	0.742	0.708	0.751	0.697
		0.727	0.766	0.727	0.775	0.727
		0.674	0.707	0.674	0.714	0.674
	MSP	0.675	0.714	0.683	0.728	0.673
		0.707	0.742	0.707	0.756	0.707
		0.636	0.675	0.636	0.685	0.636
	MI	0.208	0.525	0.158	0.530	0.166
		-0.001	0.539	-0.002	0.545	0.005
		0.054	0.492	0.061	0.493	0.049

Table 10: PRR on TriviaQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.330	0.369	0.991	0.442	0.991
		0.052	-0.052	0.416	0.387	0.485
		0.000	0.000	0.000	0.000	0.000
	CoCoA	0.334	0.311	0.987	0.267	0.987
		0.299	0.321	0.209	0.249	0.315
		0.000	0.000	0.000	0.000	0.000
	Semantic energy	0.306	0.289	0.992	0.287	0.988
		0.296	0.209	0.558	0.284	0.479
		0.000	0.000	0.000	0.000	0.000
	Semantic entropy	0.378	0.293	0.988	0.266	0.989
		-0.025	0.284	0.350	0.249	0.357
		0.000	0.000	0.000	0.000	0.000
	SAR	0.267	0.507	0.992	0.483	0.990
		0.387	0.191	0.505	0.191	0.341
		0.000	0.000	0.000	0.000	0.000
	MSP	0.470	0.479	0.984	0.482	0.988
		0.007	-0.046	0.102	-0.015	0.368
		0.000	0.000	0.000	0.000	0.000
	MI	0.192	0.502	0.990	0.496	0.998
		-0.344	0.417	0.385	0.578	0.872
		0.000	0.000	0.000	0.000	0.000
Qwen3-14B	all UQ	0.499	0.577	0.967	0.534	0.973
		0.641	0.777	0.264	0.803	0.618
		0.000	0.000	0.000	0.000	0.000
	CoCoA	0.476	0.606	0.963	0.520	0.964
		0.531	0.721	0.432	0.721	0.448
		0.000	0.000	0.000	0.000	0.000
	Semantic energy	0.468	0.489	0.971	0.436	0.972
		0.597	0.629	0.554	0.553	0.517
		0.000	0.000	0.000	0.000	0.000
	Semantic entropy	0.086	0.533	0.968	-0.545	0.956
		-0.045	0.721	-0.045	-0.564	-0.045
		0.000	0.000	0.000	0.000	0.000
	SAR	0.684	0.644	0.984	-0.319	0.986
		0.780	0.727	0.775	-0.287	0.780
		0.000	0.000	0.000	0.000	0.000
	MSP	0.444	0.623	0.951	0.645	0.971
		0.324	0.881	0.179	0.881	0.418
		0.000	0.000	0.000	0.000	0.000
	MI	-0.160	0.576	0.946	0.576	0.945
		-0.205	0.716	-0.363	0.716	-0.205
		0.000	0.000	0.000	0.000	0.000
Spectrum-Llama-3.1-8B	all UQ	0.539	0.296	0.983	0.278	0.982
		0.058	0.179	0.287	0.111	0.243
		0.000	0.000	0.000	0.000	0.000
	CoCoA	0.357	0.277	0.982	0.229	0.984
		0.360	0.214	0.125	0.089	0.164
		0.000	0.000	0.000	0.000	0.000
	Semantic energy	0.341	0.333	0.978	0.425	0.989
		0.648	0.388	-0.030	0.758	0.648
		0.000	0.000	0.000	0.000	0.000
	Semantic entropy	0.424	0.253	0.980	0.455	0.990
		0.530	0.245	0.283	0.474	0.548
		0.000	0.000	0.000	0.000	0.000
	SAR	0.301	0.367	0.984	0.379	0.980
		0.184	0.123	0.312	0.245	0.240
		0.000	0.000	0.000	0.000	0.000
	MSP	0.470	0.340	0.981	0.187	0.983
		-0.217	-0.203	0.137	-0.346	0.224
		0.000	0.000	0.000	0.000	0.000
	MI	0.385	0.433	0.984	0.181	0.988
		0.403	0.304	0.124	0.119	0.421
		0.000	0.000	0.000	0.000	0.000
Spectrum-Qwen3-14B	all UQ	0.695	0.528	0.981	0.599	0.972
		0.758	0.589	0.772	0.481	0.321
		0.000	0.000	0.000	0.000	0.000
	CoCoA	0.572	0.470	0.972	0.214	0.976
		0.712	0.506	0.559	-0.147	0.687
		0.000	0.000	0.000	0.000	0.000
	Semantic energy	0.341	0.183	0.971	-0.305	0.962
		0.073	0.258	0.585	-0.134	0.242
		0.000	0.000	0.000	0.000	0.000
	Semantic entropy	-0.005	0.269	0.978	0.267	0.947
		-0.288	0.430	0.675	0.227	0.229
		0.000	0.000	0.000	0.000	0.000
	SAR	0.381	0.136	0.982	0.212	0.948
		0.415	0.172	0.733	0.107	-0.084
		0.000	0.000	0.000	0.000	0.000
	MSP	0.236	0.464	0.965	0.161	0.961
		0.254	0.729	0.363	0.028	0.270
		0.000	0.000	0.000	0.000	0.000
	MI	0.318	0.376	0.973	0.084	0.974
		0.451	0.473	0.504	0.211	0.577
		0.000	0.000	0.000	0.000	0.000
A ² Search-7B	all UQ	0.491	0.388	0.923	0.336	0.908
		-0.253	-0.332	0.521	0.054	0.195
		0.886	0.751	0.951	0.636	0.962
	CoCoA	0.287	0.013	0.919	-0.287	0.891
		0.123	-0.584	0.418	-0.567	-0.101
		0.935	0.229	0.944	-0.288	0.943
	Semantic energy	0.463	0.561	0.917	0.630	0.883
		-0.123	0.043	0.521	0.145	-0.123
		0.933	0.960	0.933	0.976	0.933
	Semantic entropy	0.491	0.446	0.851	0.420	0.830
		-0.641	-0.056	0.337	-0.080	-0.641
		0.917	0.953	0.917	0.908	0.919
	SAR	0.683	0.670	0.910	0.330	0.914
		0.474	0.812	0.358	0.075	0.252
		0.928	0.952	0.960	0.748	0.965
	MSP	0.499	0.638	0.929	0.503	0.910
		0.147	0.601	0.806	0.601	0.169
		0.963	0.987	0.963	0.984	0.963
	MI	0.365	0.479	0.704	0.474	0.704
		-0.322	0.530	-0.322	0.530	-0.322
		0.774	0.770	0.780	0.814	0.780

Table 11: PRR on NCQA (test split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.788	0.788	0.790	0.786	0.788
		0.827	0.828	0.829	0.824	0.827
		0.729	0.730	0.725	0.722	0.728
	CoCoA	0.792	0.790	0.795	0.788	0.792
		0.834	0.833	0.836	0.834	0.835
		0.736	0.735	0.737	0.735	0.737
	Semantic energy	0.777	0.777	0.779	0.774	0.776
		0.816	0.815	0.816	0.814	0.816
		0.726	0.726	0.726	0.721	0.726
	Semantic entropy	0.780	0.780	0.782	0.766	0.780
		0.818	0.818	0.818	0.813	0.818
		0.732	0.730	0.732	0.725	0.732
	SAR	0.776	0.635	0.664	0.626	0.784
		0.342	0.352	0.781	0.382	0.784
		0.398	0.442	0.730	0.508	0.709
	MSP	0.729	0.651	0.656	0.666	0.730
		0.213	0.543	0.697	0.652	0.707
		0.295	0.500	0.417	0.611	0.636
	MI	0.576	0.575	0.590	0.574	0.577
		0.577	0.564	0.578	0.578	0.576
		0.568	0.566	0.562	0.563	0.568
Qwen3-14B	all UQ	0.758	0.759	0.763	0.758	0.756
		0.800	0.808	0.808	0.806	0.804
		0.694	0.691	0.694	0.690	0.694
	CoCoA	0.761	0.763	0.763	0.749	0.760
		0.813	0.810	0.813	0.812	0.813
		0.694	0.686	0.694	0.688	0.694
	Semantic energy	0.728	0.738	0.742	0.738	0.728
		0.781	0.785	0.785	0.781	0.781
		0.661	0.653	0.668	0.638	0.661
	Semantic entropy	0.740	0.747	0.748	0.744	0.740
		0.789	0.799	0.789	0.798	0.789
		0.678	0.676	0.678	0.668	0.678
	SAR	0.770	0.639	0.597	0.652	0.634
		0.207	0.588	0.619	0.482	0.664
		0.320	0.484	0.383	0.429	0.602
	MSP	0.642	0.527	0.581	0.475	0.358
		0.270	0.407	0.323	0.446	0.269
		0.383	0.485	0.626	0.572	0.382
	MI	0.503	0.507	0.520	0.507	0.502
		0.500	0.504	0.502	0.497	0.500
		0.503	0.491	0.503	0.513	0.503
Spectrum-Llama-3.1-8B	all UQ	0.748	0.746	0.749	0.742	0.749
		0.771	0.772	0.768	0.765	0.770
		0.715	0.711	0.711	0.700	0.714
	CoCoA	0.736	0.735	0.737	0.723	0.736
		0.756	0.757	0.757	0.747	0.757
		0.708	0.702	0.708	0.693	0.708
	Semantic energy	0.728	0.727	0.730	0.702	0.728
		0.753	0.749	0.754	0.734	0.754
		0.693	0.693	0.695	0.662	0.695
	Semantic entropy	0.730	0.730	0.731	0.688	0.730
		0.756	0.755	0.756	0.720	0.756
		0.695	0.695	0.695	0.650	0.695
	SAR	0.741	0.740	0.742	0.732	0.741
		0.765	0.764	0.765	0.760	0.765
		0.708	0.705	0.708	0.696	0.708
	MSP	0.706	0.706	0.706	0.704	0.706
		0.724	0.724	0.724	0.725	0.724
		0.682	0.677	0.682	0.680	0.682
	MI	0.571	0.581	0.575	0.550	0.578
		0.571	0.577	0.558	0.560	0.571
		0.569	0.581	0.562	0.545	0.572
Spectrum-Qwen3-14B	all UQ	0.798	0.798	0.801	0.792	0.796
		0.817	0.817	0.820	0.811	0.816
		0.773	0.772	0.773	0.767	0.771
	CoCoA	0.777	0.776	0.776	0.765	0.777
		0.794	0.794	0.794	0.779	0.794
		0.753	0.752	0.753	0.740	0.753
	Semantic energy	0.771	0.772	0.771	0.771	0.771
		0.791	0.792	0.792	0.783	0.792
		0.741	0.743	0.743	0.731	0.743
	Semantic entropy	0.780	0.780	0.781	0.775	0.780
		0.800	0.800	0.801	0.796	0.801
		0.753	0.753	0.753	0.745	0.753
	SAR	0.770	0.771	0.771	0.756	0.770
		0.795	0.795	0.795	0.781	0.795
		0.736	0.735	0.736	0.726	0.736
	MSP	0.749	0.749	0.748	0.738	0.748
		0.765	0.763	0.763	0.756	0.763
		0.728	0.727	0.727	0.717	0.727
	MI	0.580	0.588	0.587	0.588	0.578
		0.576	0.578	0.576	0.555	0.575
		0.585	0.586	0.582	0.559	0.583
A2Search-7B	all UQ	0.782	0.777	0.785	0.775	0.782
		0.823	0.819	0.824	0.813	0.820
		0.727	0.721	0.727	0.711	0.725
	CoCoA	0.773	0.774	0.777	0.734	0.773
		0.808	0.807	0.808	0.768	0.808
		0.730	0.730	0.730	0.703	0.730
	Semantic energy	0.754	0.758	0.765	0.745	0.754
		0.798	0.800	0.801	0.777	0.798
		0.703	0.706	0.711	0.683	0.703
	Semantic entropy	0.766	0.769	0.773	0.761	0.765
		0.807	0.807	0.807	0.804	0.807
		0.715	0.715	0.714	0.711	0.714
	SAR	0.760	0.721	0.662	0.720	0.760
		0.785	0.734	0.760	0.275	0.788
		0.704	0.685	0.683	0.348	0.706
	MSP	0.640	0.502	0.546	0.586	0.361
		0.279	0.452	0.266	0.267	0.322
		0.335	0.481	0.507	0.339	0.374
	MI	0.506	0.547	0.525	0.547	0.501
		0.501	0.557	0.502	0.440	0.502
		0.513	0.519	0.498	0.474	0.506

Table 12: AUROC (error prediction) on AmbigQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best AUROC for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.886	0.887	0.887	0.888	0.886
		0.894	0.894	0.893	0.894	0.893
		0.875	0.876	0.876	0.877	0.875
	CoCoA	0.874	0.874	0.874	0.875	0.874
		0.881	0.881	0.881	0.882	0.881
		0.863	0.863	0.863	0.864	0.863
	Semantic energy	0.865	0.866	0.866	0.867	0.865
		0.872	0.873	0.872	0.874	0.872
		0.853	0.855	0.853	0.856	0.853
	Semantic entropy	0.868	0.870	0.867	0.871	0.868
		0.874	0.876	0.874	0.876	0.874
		0.858	0.860	0.858	0.861	0.858
	SAR	0.873	0.873	0.873	0.874	0.873
		0.880	0.880	0.880	0.881	0.880
		0.861	0.862	0.861	0.863	0.861
	MSP	0.847	0.850	0.847	0.853	0.847
		0.855	0.857	0.855	0.860	0.855
		0.835	0.838	0.835	0.841	0.835
	MI	0.838	0.845	0.839	0.848	0.838
		0.842	0.847	0.842	0.850	0.842
		0.829	0.837	0.829	0.840	0.829
Qwen3-14B	all UQ	0.860	0.864	0.860	0.872	0.859
		0.862	0.865	0.862	0.873	0.862
		0.854	0.859	0.854	0.866	0.854
	CoCoA	0.840	0.846	0.840	0.854	0.839
		0.842	0.848	0.842	0.856	0.842
		0.833	0.840	0.833	0.846	0.833
	Semantic energy	0.846	0.852	0.847	0.861	0.846
		0.849	0.854	0.849	0.861	0.849
		0.840	0.846	0.840	0.854	0.840
	Semantic entropy	0.814	0.846	0.821	0.848	0.814
		0.816	0.847	0.816	0.849	0.816
		0.809	0.840	0.809	0.842	0.809
	SAR	0.843	0.854	0.845	0.861	0.843
		0.845	0.854	0.845	0.861	0.844
		0.838	0.848	0.839	0.856	0.837
	MSP	0.828	0.836	0.828	0.843	0.828
		0.830	0.838	0.830	0.846	0.830
		0.821	0.829	0.821	0.835	0.821
	MI	0.542	0.735	0.588	0.735	0.541
		0.551	0.733	0.551	0.733	0.551
		0.534	0.723	0.532	0.723	0.535
Spectrum-Llama-3.1-8B	all UQ	0.876	0.877	0.877	0.879	0.877
		0.886	0.886	0.886	0.888	0.885
		0.864	0.865	0.865	0.866	0.865
	CoCoA	0.859	0.861	0.860	0.863	0.859
		0.868	0.870	0.868	0.872	0.868
		0.846	0.848	0.846	0.849	0.846
	Semantic energy	0.857	0.861	0.858	0.863	0.857
		0.861	0.865	0.861	0.868	0.861
		0.849	0.852	0.849	0.854	0.849
	Semantic entropy	0.860	0.864	0.861	0.867	0.860
		0.866	0.870	0.866	0.873	0.866
		0.851	0.855	0.851	0.856	0.851
	SAR	0.864	0.865	0.864	0.867	0.864
		0.872	0.873	0.872	0.875	0.872
		0.851	0.852	0.851	0.854	0.851
	MSP	0.820	0.826	0.821	0.833	0.820
		0.828	0.833	0.828	0.842	0.828
		0.807	0.812	0.807	0.818	0.807
	MI	0.824	0.834	0.820	0.839	0.824
		0.826	0.835	0.826	0.842	0.826
		0.817	0.826	0.811	0.829	0.818
Spectrum-Qwen3-14B	all UQ	0.898	0.898	0.899	0.900	0.898
		0.900	0.901	0.902	0.903	0.901
		0.892	0.892	0.892	0.893	0.892
	CoCoA	0.884	0.885	0.884	0.887	0.884
		0.888	0.889	0.888	0.891	0.888
		0.876	0.877	0.876	0.878	0.876
	Semantic energy	0.877	0.879	0.877	0.880	0.877
		0.880	0.882	0.880	0.884	0.880
		0.870	0.871	0.870	0.873	0.870
	Semantic entropy	0.882	0.883	0.882	0.886	0.882
		0.885	0.887	0.885	0.890	0.885
		0.875	0.876	0.875	0.878	0.875
	SAR	0.886	0.887	0.886	0.888	0.886
		0.888	0.889	0.888	0.891	0.888
		0.881	0.882	0.881	0.882	0.881
	MSP	0.853	0.856	0.854	0.861	0.853
		0.858	0.861	0.858	0.866	0.858
		0.845	0.848	0.845	0.850	0.845
	MI	0.855	0.861	0.851	0.863	0.855
		0.846	0.858	0.842	0.861	0.847
		0.851	0.859	0.849	0.860	0.851
A2Search-7B	all UQ	0.845	0.848	0.846	0.855	0.844
		0.855	0.858	0.855	0.865	0.855
		0.832	0.835	0.832	0.841	0.831
	CoCoA	0.821	0.831	0.822	0.836	0.821
		0.833	0.843	0.833	0.848	0.833
		0.806	0.817	0.806	0.819	0.806
	Semantic energy	0.838	0.844	0.839	0.850	0.838
		0.852	0.856	0.852	0.863	0.852
		0.823	0.829	0.823	0.834	0.823
	Semantic entropy	0.815	0.838	0.820	0.841	0.815
		0.825	0.849	0.825	0.852	0.825
		0.802	0.823	0.802	0.825	0.802
	SAR	0.830	0.840	0.831	0.845	0.830
		0.840	0.849	0.840	0.854	0.840
		0.818	0.827	0.818	0.832	0.818
	MSP	0.810	0.821	0.812	0.828	0.810
		0.824	0.834	0.824	0.841	0.824
		0.794	0.804	0.794	0.809	0.794
	MI	0.552	0.717	0.588	0.721	0.553
		0.554	0.720	0.554	0.725	0.554
		0.552	0.706	0.553	0.708	0.551

Table 13: AUROC (error prediction) on TriviaQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best AUROC for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.615	0.599	0.991	0.649	0.578
		0.533	0.578	0.756	0.756	0.756
		0.500	0.500	0.500	0.500	0.500
	CoCoA	0.598	0.584	0.988	0.542	0.610
		0.622	0.622	0.644	0.600	0.644
		0.500	0.500	0.500	0.500	0.500
	Semantic energy	0.569	0.591	0.992	0.569	0.508
		0.622	0.644	0.778	0.578	0.667
		0.500	0.500	0.500	0.500	0.500
	Semantic entropy	0.620	0.581	0.989	0.547	0.490
		0.556	0.622	0.689	0.622	0.711
		0.500	0.500	0.500	0.500	0.500
	SAR	0.609	0.652	0.992	0.632	0.605
		0.778	0.622	0.778	0.644	0.733
		0.500	0.500	0.500	0.500	0.500
	MSP	0.636	0.639	0.985	0.637	0.590
		0.511	0.556	0.578	0.578	0.689
		0.500	0.500	0.500	0.500	0.500
	MI	0.529	0.665	0.991	0.646	0.618
		0.289	0.778	0.733	0.800	0.956
		0.500	0.500	0.500	0.500	0.500
Qwen3-14B	all UQ	0.667	0.723	0.970	0.704	0.655
		0.775	0.875	0.725	0.900	0.762
		0.500	0.500	0.500	0.500	0.500
	CoCoA	0.641	0.761	0.965	0.741	0.628
		0.713	0.887	0.675	0.887	0.688
		0.500	0.500	0.500	0.500	0.500
	Semantic energy	0.677	0.711	0.974	0.678	0.644
		0.794	0.825	0.756	0.875	0.756
		0.500	0.500	0.500	0.500	0.500
	Semantic entropy	0.598	0.745	0.959	0.745	0.402
		0.625	0.887	0.625	0.362	0.625
		0.500	0.500	0.500	0.500	0.500
	SAR	0.794	0.776	0.984	0.768	0.730
		0.869	0.850	0.856	0.850	0.869
		0.500	0.500	0.500	0.500	0.500
	MSP	0.646	0.746	0.954	0.712	0.681
		0.675	0.938	0.575	0.938	0.750
		0.500	0.500	0.500	0.500	0.500
	MI	0.558	0.743	0.952	0.743	0.556
		0.562	0.838	0.438	0.838	0.562
		0.500	0.500	0.500	0.500	0.500
Spectrum-Llama-3.1-8B	all UQ	0.672	0.578	0.983	0.594	0.599
		0.446	0.589	0.643	0.607	0.625
		0.500	0.500	0.500	0.500	0.500
	CoCoA	0.579	0.559	0.983	0.571	0.505
		0.643	0.589	0.643	0.536	0.679
		0.500	0.500	0.500	0.500	0.500
	Semantic energy	0.591	0.594	0.979	0.601	0.591
		0.795	0.732	0.545	0.839	0.795
		0.500	0.500	0.500	0.500	0.500
	Semantic entropy	0.664	0.610	0.981	0.647	0.654
		0.786	0.643	0.589	0.714	0.804
		0.500	0.500	0.500	0.500	0.500
	SAR	0.556	0.631	0.985	0.679	0.599
		0.518	0.607	0.679	0.750	0.589
		0.500	0.500	0.500	0.500	0.500
	MSP	0.662	0.630	0.982	0.549	0.500
		0.375	0.429	0.607	0.304	0.661
		0.500	0.500	0.500	0.500	0.500
	MI	0.632	0.647	0.985	0.557	0.642
		0.750	0.643	0.679	0.607	0.768
		0.500	0.500	0.500	0.500	0.500
Spectrum-Qwen3-14B	all UQ	0.752	0.680	0.982	0.745	0.678
		0.815	0.827	0.815	0.765	0.741
		0.500	0.500	0.500	0.500	0.500
	CoCoA	0.681	0.614	0.973	0.542	0.663
		0.790	0.654	0.728	0.494	0.765
		0.500	0.500	0.500	0.500	0.500
	Semantic energy	0.540	0.548	0.972	0.559	0.639
		0.432	0.617	0.716	0.630	0.642
		0.500	0.500	0.500	0.500	0.500
	Semantic entropy	0.435	0.570	0.979	0.617	0.543
		0.333	0.679	0.790	0.642	0.481
		0.500	0.500	0.500	0.500	0.500
	SAR	0.576	0.508	0.983	0.570	0.462
		0.605	0.630	0.827	0.580	0.506
		0.500	0.500	0.500	0.500	0.500
	MSP	0.586	0.632	0.967	0.602	0.557
		0.617	0.827	0.667	0.630	0.630
		0.500	0.500	0.500	0.500	0.500
	MI	0.565	0.636	0.974	0.597	0.581
		0.698	0.753	0.722	0.741	0.735
		0.500	0.500	0.500	0.500	0.500
A ² Search-7B	all UQ	0.662	0.604	0.931	0.649	0.667
		0.321	0.284	0.679	0.531	0.481
		0.906	0.809	0.954	0.818	0.965
	CoCoA	0.608	0.515	0.928	0.484	0.602
		0.469	0.235	0.599	0.222	0.401
		0.941	0.684	0.950	0.667	0.947
	Semantic energy	0.652	0.712	0.928	0.739	0.630
		0.358	0.444	0.679	0.691	0.358
		0.941	0.963	0.941	0.932	0.941
	Semantic entropy	0.620	0.688	0.897	0.689	0.624
		0.333	0.457	0.667	0.432	0.333
		0.890	0.958	0.890	0.967	0.895
	SAR	0.765	0.766	0.918	0.699	0.748
		0.611	0.877	0.580	0.815	0.556
		0.932	0.956	0.963	0.928	0.967
	MSP	0.690	0.768	0.938	0.722	0.683
		0.457	0.716	0.889	0.716	0.481
		0.965	0.987	0.965	0.985	0.965
	MI	0.598	0.668	0.837	0.650	0.600
		0.500	0.790	0.500	0.790	0.500
		0.743	0.803	0.750	0.770	0.750

Table 14: AUROC (error prediction) on NCQA (test split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best AUROC for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.641	0.628	0.649	0.626	0.623
		0.670	0.635	0.673	0.622	0.670
		0.602	0.588	0.595	0.570	0.596
	CoCoA	0.644	0.644	0.648	0.628	0.629
		0.683	0.675	0.684	0.655	0.683
		0.591	0.594	0.592	0.585	0.592
	Semantic energy	0.618	0.617	0.620	0.612	0.602
		0.651	0.644	0.651	0.649	0.651
		0.571	0.572	0.571	0.565	0.571
	Semantic entropy	0.612	0.614	0.615	0.602	0.591
		0.651	0.648	0.651	0.633	0.651
		0.558	0.563	0.557	0.555	0.557
	SAR	0.548	0.333	0.418	0.382	0.346
		0.547	0.011	0.462	0.376	0.373
		0.420	0.098	0.053	0.382	0.218
	MSP	0.413	0.273	0.351	0.185	0.346
		0.319	-0.328	-0.313	0.138	0.307
		0.265	-0.190	0.489	0.148	0.328
	MI	0.228	0.197	0.239	0.184	0.225
		0.180	0.147	0.161	0.140	0.173
		0.219	0.216	0.223	0.208	0.210
Qwen3-14B	all UQ	0.480	0.475	0.516	0.446	0.480
		0.501	0.512	0.541	0.479	0.508
		0.438	0.419	0.453	0.383	0.430
	CoCoA	0.535	0.519	0.540	0.477	0.519
		0.581	0.567	0.581	0.515	0.581
		0.474	0.442	0.474	0.404	0.474
	Semantic energy	0.463	0.442	0.485	0.432	0.467
		0.521	0.485	0.502	0.467	0.521
		0.399	0.366	0.391	0.355	0.400
	Semantic entropy	0.453	0.461	0.471	0.427	0.449
		0.484	0.500	0.485	0.458	0.484
		0.393	0.387	0.390	0.363	0.388
	SAR	0.547	0.368	0.417	0.257	0.434
		0.584	0.392	0.566	0.308	0.598
		0.451	0.305	0.474	0.152	0.456
	MSP	0.217	0.213	0.320	0.222	0.380
		0.221	0.071	0.137	0.150	0.526
		0.212	0.053	0.181	0.032	0.410
	MI	0.029	0.054	0.104	0.054	0.099
		-0.030	0.015	-0.043	0.015	-0.012
		-0.004	0.044	0.003	0.041	-0.004
Spectrum-Llama-3.1-8B	all UQ	0.634	0.632	0.635	0.628	0.611
		0.652	0.647	0.654	0.633	0.645
		0.606	0.607	0.606	0.603	0.603
	CoCoA	0.567	0.569	0.576	0.564	0.548
		0.577	0.575	0.588	0.565	0.579
		0.550	0.552	0.552	0.555	0.549
	Semantic energy	0.554	0.552	0.559	0.512	0.505
		0.565	0.554	0.571	0.516	0.569
		0.522	0.531	0.528	0.495	0.526
	Semantic entropy	0.555	0.549	0.559	0.509	0.511
		0.577	0.557	0.577	0.510	0.577
		0.523	0.520	0.525	0.505	0.525
	SAR	0.619	0.618	0.621	0.612	0.606
		0.632	0.628	0.632	0.618	0.632
		0.596	0.599	0.596	0.598	0.596
	MSP	0.524	0.528	0.525	0.515	0.502
		0.530	0.532	0.531	0.504	0.531
		0.509	0.514	0.509	0.504	0.509
	MI	0.191	0.257	0.236	0.235	0.207
		0.218	0.253	0.199	0.226	0.207
		0.213	0.254	0.189	0.234	0.225
Spectrum-Qwen3-14B	all UQ	0.711	0.710	0.717	0.705	0.695
		0.734	0.733	0.737	0.720	0.735
		0.676	0.672	0.679	0.657	0.680
	CoCoA	0.671	0.670	0.672	0.666	0.647
		0.691	0.691	0.692	0.689	0.692
		0.639	0.636	0.641	0.632	0.640
	Semantic energy	0.645	0.645	0.649	0.628	0.625
		0.674	0.675	0.676	0.656	0.676
		0.600	0.600	0.599	0.589	0.599
	Semantic entropy	0.655	0.656	0.659	0.652	0.632
		0.684	0.684	0.685	0.678	0.685
		0.617	0.612	0.618	0.611	0.618
	SAR	0.684	0.685	0.688	0.678	0.672
		0.713	0.713	0.713	0.707	0.713
		0.640	0.639	0.640	0.631	0.640
	MSP	0.627	0.626	0.627	0.620	0.603
		0.640	0.639	0.640	0.633	0.640
		0.604	0.600	0.604	0.595	0.604
	MI	0.293	0.260	0.285	0.254	0.276
		0.208	0.216	0.204	0.208	0.180
		0.247	0.288	0.246	0.280	0.248
A2Search-7B	all UQ	0.551	0.542	0.568	0.523	0.537
		0.570	0.549	0.572	0.530	0.566
		0.533	0.499	0.562	0.483	0.528
	CoCoA	0.557	0.560	0.564	0.527	0.530
		0.570	0.570	0.570	0.530	0.570
		0.546	0.547	0.546	0.516	0.546
	Semantic energy	0.501	0.513	0.523	0.503	0.506
		0.536	0.542	0.536	0.524	0.536
		0.479	0.480	0.478	0.466	0.478
	Semantic entropy	0.523	0.533	0.543	0.523	0.512
		0.544	0.544	0.544	0.536	0.543
		0.506	0.509	0.502	0.495	0.502
	SAR	0.099	0.382	0.399	0.296	0.106
		-0.291	0.308	-0.118	0.296	0.065
		-0.239	0.237	-0.042	0.263	0.125
	MSP	0.457	0.008	-0.052	0.011	0.338
		-0.333	-0.237	-0.366	0.004	-0.355
		-0.234	-0.243	0.030	-0.009	-0.366
	MI	0.013	0.110	0.063	0.112	0.059
		0.010	0.135	-0.004	0.134	-0.015
		-0.005	0.032	-0.001	0.042	-0.017

Table 15: PRR on AmbigQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.878	0.880	0.879	0.879	0.870
		0.890	0.892	0.890	0.892	0.891
		0.861	0.863	0.860	0.862	0.861
	CoCoA	0.850	0.850	0.850	0.850	0.842
		0.860	0.861	0.860	0.861	0.860
		0.834	0.834	0.834	0.834	0.834
	Semantic energy	0.835	0.837	0.835	0.838	0.826
		0.848	0.849	0.848	0.851	0.848
		0.815	0.817	0.815	0.817	0.815
	Semantic entropy	0.840	0.844	0.842	0.846	0.834
		0.854	0.860	0.854	0.861	0.854
		0.823	0.823	0.823	0.823	0.824
	SAR	0.861	0.862	0.861	0.861	0.851
		0.874	0.875	0.874	0.874	0.874
		0.842	0.843	0.842	0.841	0.842
	MSP	0.808	0.811	0.808	0.815	0.797
		0.815	0.819	0.815	0.824	0.815
		0.793	0.796	0.793	0.797	0.793
	MI	0.809	0.823	0.813	0.826	0.798
		0.816	0.834	0.817	0.836	0.817
		0.788	0.804	0.787	0.807	0.787
Qwen3-14B	all UQ	0.790	0.801	0.793	0.819	0.782
		0.810	0.816	0.807	0.829	0.808
		0.765	0.774	0.764	0.800	0.759
	CoCoA	0.752	0.769	0.752	0.779	0.745
		0.767	0.775	0.767	0.782	0.767
		0.731	0.753	0.731	0.765	0.731
	Semantic energy	0.753	0.773	0.760	0.789	0.751
		0.780	0.790	0.779	0.797	0.781
		0.719	0.744	0.721	0.771	0.719
	Semantic entropy	0.652	0.758	0.679	0.761	0.676
		0.670	0.757	0.678	0.760	0.678
		0.640	0.748	0.626	0.750	0.626
	SAR	0.744	0.778	0.752	0.783	0.732
		0.758	0.777	0.761	0.779	0.758
		0.723	0.771	0.728	0.776	0.723
	MSP	0.735	0.757	0.739	0.765	0.728
		0.751	0.764	0.751	0.770	0.751
		0.711	0.740	0.711	0.749	0.711
	MI	0.211	0.577	0.192	0.577	0.185
		-0.038	0.565	-0.038	0.566	-0.038
		0.006	0.567	-0.006	0.567	0.006
Spectrum-Llama-3.1-8B	all UQ	0.855	0.856	0.855	0.859	0.847
		0.874	0.874	0.873	0.877	0.874
		0.831	0.833	0.831	0.837	0.831
	CoCoA	0.809	0.813	0.809	0.816	0.798
		0.823	0.827	0.823	0.829	0.823
		0.790	0.793	0.790	0.796	0.790
	Semantic energy	0.812	0.819	0.814	0.823	0.805
		0.827	0.833	0.827	0.839	0.827
		0.792	0.798	0.792	0.801	0.792
	Semantic entropy	0.819	0.828	0.820	0.831	0.811
		0.835	0.844	0.836	0.846	0.835
		0.798	0.807	0.798	0.808	0.798
	SAR	0.834	0.839	0.834	0.841	0.824
		0.848	0.854	0.848	0.856	0.848
		0.815	0.819	0.815	0.821	0.815
	MSP	0.739	0.749	0.740	0.764	0.730
		0.750	0.760	0.750	0.778	0.750
		0.721	0.731	0.721	0.740	0.721
	MI	0.779	0.797	0.761	0.797	0.749
		0.787	0.808	0.787	0.808	0.787
		0.761	0.780	0.702	0.775	0.761
Spectrum-Qwen3-14B	all UQ	0.877	0.878	0.878	0.880	0.870
		0.888	0.889	0.888	0.892	0.888
		0.861	0.861	0.860	0.861	0.861
	CoCoA	0.846	0.849	0.847	0.849	0.836
		0.855	0.860	0.855	0.861	0.855
		0.830	0.831	0.830	0.831	0.830
	Semantic energy	0.835	0.838	0.836	0.842	0.828
		0.850	0.853	0.850	0.857	0.850
		0.815	0.816	0.815	0.820	0.815
	Semantic entropy	0.841	0.849	0.843	0.851	0.838
		0.856	0.861	0.857	0.864	0.856
		0.821	0.830	0.821	0.831	0.821
	SAR	0.862	0.864	0.862	0.864	0.853
		0.871	0.873	0.871	0.874	0.871
		0.846	0.847	0.846	0.847	0.846
	MSP	0.793	0.800	0.796	0.804	0.786
		0.800	0.809	0.800	0.817	0.800
		0.778	0.782	0.778	0.784	0.778
	MI	0.814	0.827	0.785	0.823	0.777
		0.784	0.830	0.787	0.830	0.793
		0.773	0.810	0.771	0.805	0.762
A2Search-7B	all UQ	0.736	0.750	0.741	0.769	0.732
		0.758	0.768	0.753	0.784	0.756
		0.708	0.725	0.709	0.745	0.707
	CoCoA	0.672	0.719	0.678	0.728	0.672
		0.694	0.735	0.694	0.744	0.694
		0.643	0.693	0.643	0.699	0.643
	Semantic energy	0.710	0.731	0.718	0.749	0.707
		0.734	0.753	0.737	0.768	0.726
		0.681	0.700	0.677	0.719	0.674
	Semantic entropy	0.655	0.732	0.662	0.736	0.655
		0.671	0.741	0.675	0.744	0.675
		0.627	0.714	0.630	0.717	0.630
	SAR	0.696	0.736	0.695	0.745	0.692
		0.712	0.748	0.712	0.756	0.712
		0.673	0.716	0.673	0.723	0.673
	MSP	0.651	0.693	0.663	0.712	0.656
		0.676	0.712	0.676	0.727	0.676
		0.620	0.665	0.620	0.682	0.620
	MI	0.222	0.486	0.155	0.491	0.156
		-0.020	0.480	0.004	0.487	-0.004
		0.027	0.469	0.028	0.470	0.024

Table 16: PRR on TriviaQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.582	0.829	0.703	0.798	0.552
		1.000	0.795	-0.480	0.633	0.795
		0.478	0.828	0.864	0.833	0.647
	CoCoA	0.513	0.506	0.600	0.392	0.498
		1.000	0.542	0.213	0.334	-1.000
		0.419	0.476	0.724	0.395	0.620
	Semantic energy	0.462	0.496	0.597	0.491	0.520
		-0.752	0.076	1.000	-0.480	-0.752
		0.720	0.600	0.696	0.756	0.639
	Semantic entropy	0.069	0.573	0.386	0.566	0.430
		0.633	0.795	-0.480	0.633	-0.480
		-0.023	0.523	0.468	0.547	0.523
	SAR	0.375	0.619	0.650	0.080	0.628
		-1.000	-0.261	0.868	-0.480	-0.752
		0.734	0.813	0.768	0.198	0.774
	MSP	0.763	0.738	0.680	0.684	0.633
		0.936	0.542	0.868	0.717	0.936
		0.712	0.761	0.805	0.674	0.747
	MI	0.425	0.714	0.529	0.707	0.376
		0.717	0.633	-0.261	0.542	0.717
		0.385	0.735	0.648	0.741	0.378
Qwen3-14B	all UQ	0.735	0.919	0.714	0.925	0.626
		0.000	0.000	0.000	0.000	0.000
		0.718	0.922	0.758	0.922	0.722
	CoCoA	0.900	0.489	0.683	0.399	0.758
		0.000	0.000	0.000	0.000	0.000
		0.913	0.538	0.891	0.450	0.924
	Semantic energy	0.921	0.767	0.638	0.636	0.495
		0.000	0.000	0.000	0.000	0.000
		0.913	0.777	0.848	0.634	0.702
	Semantic entropy	0.511	0.245	0.495	0.245	0.397
		0.000	0.000	0.000	0.000	0.000
		-0.110	0.280	-0.110	0.280	-0.110
	SAR	0.806	0.611	0.573	0.560	0.558
		0.000	0.000	0.000	0.000	0.000
		0.832	0.614	0.322	0.551	0.775
	MSP	0.892	0.922	0.867	0.761	0.734
		0.000	0.000	0.000	0.000	0.000
		0.891	0.921	0.862	0.753	0.891
	MI	0.472	0.583	0.353	0.550	0.353
		0.000	0.000	0.000	0.000	0.000
		-0.279	0.599	-0.279	0.579	-0.279
Spectrum-Llama-3.1-8B	all UQ	0.434	0.338	0.375	0.646	0.437
		0.076	-1.000	-0.480	1.000	-0.080
		0.491	0.462	0.480	0.546	0.547
	CoCoA	0.408	0.487	0.239	0.524	0.385
		-0.261	0.868	-0.080	0.868	-0.261
		0.518	0.392	0.311	0.468	0.490
	Semantic energy	0.482	0.559	0.419	0.540	0.347
		1.000	0.542	0.443	0.795	0.443
		0.353	0.569	0.414	0.448	0.437
	Semantic entropy	-0.099	0.424	0.597	0.423	0.166
		-1.000	0.795	0.542	0.795	0.213
		0.115	0.315	0.675	0.345	0.243
	SAR	0.788	0.767	0.644	0.726	0.716
		1.000	1.000	0.334	0.868	1.000
		0.758	0.734	0.660	0.687	0.758
	MSP	0.182	0.378	0.367	0.341	0.333
		-1.000	0.936	0.936	0.633	-1.000
		0.455	0.230	0.442	0.273	0.437
	MI	0.625	0.503	0.395	0.519	0.664
		-0.480	-1.000	-0.261	-0.480	-1.000
		0.806	0.690	0.491	0.726	0.841
Spectrum-Qwen3-14B	all UQ	0.823	0.898	0.889	0.942	0.851
		0.443	1.000	0.443	1.000	0.076
		0.928	0.881	0.936	0.945	0.932
	CoCoA	0.557	0.762	0.557	0.719	0.496
		0.795	1.000	1.000	1.000	0.795
		0.541	0.711	0.416	0.669	0.577
	Semantic energy	0.493	0.718	0.746	0.721	0.552
		-0.080	0.542	0.213	0.868	0.076
		0.594	0.762	0.842	0.703	0.626
	Semantic entropy	0.618	0.649	0.518	0.627	0.620
		0.213	0.334	-0.261	0.868	-0.752
		0.679	0.709	0.630	0.587	0.763
	SAR	0.604	0.592	0.587	0.598	0.568
		0.443	0.443	0.633	0.334	0.334
		0.637	0.634	0.620	0.683	0.681
	MSP	0.685	0.748	0.646	0.732	0.574
		0.633	1.000	0.542	1.000	0.633
		0.712	0.690	0.773	0.659	0.680
	MI	0.546	0.681	0.707	0.714	0.650
		-0.080	0.334	-0.080	0.868	-0.080
		0.757	0.782	0.788	0.681	0.788
A ² Search-7B	all UQ	0.615	0.673	0.381	0.550	0.598
		0.000	0.000	0.000	0.000	0.000
		0.641	0.643	0.528	0.557	0.808
	CoCoA	0.479	0.516	0.198	0.648	0.346
		0.000	0.000	0.000	0.000	0.000
		0.526	0.502	0.404	0.649	0.554
	Semantic energy	0.630	0.527	0.569	0.488	0.479
		0.000	0.000	0.000	0.000	0.000
		0.630	0.552	0.779	0.514	0.613
	Semantic entropy	0.475	0.441	0.341	0.277	0.337
		0.000	0.000	0.000	0.000	0.000
		0.052	0.408	0.165	0.301	0.052
	SAR	0.817	0.596	0.593	0.316	0.636
		0.000	0.000	0.000	0.000	0.000
		0.832	0.602	0.774	0.302	0.796
	MSP	0.681	0.600	0.306	0.566	0.361
		0.000	0.000	0.000	0.000	0.000
		0.666	0.595	0.514	0.553	0.603
	MI	0.424	0.417	0.246	0.398	0.268
		0.000	0.000	0.000	0.000	0.000
		-0.086	0.421	-0.086	0.397	-0.086

Table 17: PRR on NCQA (test split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.772	0.770	0.773	0.765	0.771
		0.783	0.775	0.782	0.779	0.784
		0.757	0.753	0.754	0.741	0.756
	CoCoA	0.768	0.768	0.769	0.759	0.768
		0.783	0.777	0.783	0.778	0.783
		0.750	0.750	0.751	0.742	0.750
	Semantic energy	0.752	0.751	0.752	0.749	0.752
		0.764	0.761	0.764	0.768	0.764
		0.737	0.736	0.737	0.732	0.737
	Semantic entropy	0.754	0.754	0.753	0.752	0.754
		0.769	0.769	0.768	0.771	0.768
		0.736	0.736	0.736	0.733	0.736
	SAR	0.745	0.658	0.673	0.707	0.703
		0.742	0.507	0.712	0.726	0.691
		0.700	0.569	0.572	0.691	0.637
	MSP	0.705	0.636	0.632	0.681	0.715
		0.622	0.341	0.340	0.679	0.630
		0.602	0.406	0.674	0.623	0.625
	MI	0.579	0.575	0.594	0.577	0.578
		0.579	0.553	0.568	0.592	0.575
		0.582	0.584	0.582	0.575	0.581
Qwen3-14B	all UQ	0.714	0.711	0.721	0.713	0.707
		0.720	0.723	0.729	0.725	0.717
		0.698	0.696	0.703	0.698	0.696
	CoCoA	0.718	0.716	0.721	0.705	0.718
		0.729	0.725	0.729	0.716	0.729
		0.706	0.704	0.706	0.697	0.706
	Semantic energy	0.704	0.701	0.709	0.698	0.704
		0.720	0.713	0.718	0.718	0.720
		0.687	0.685	0.688	0.679	0.687
	Semantic entropy	0.698	0.699	0.701	0.699	0.698
		0.708	0.706	0.708	0.710	0.708
		0.685	0.683	0.685	0.684	0.684
	SAR	0.741	0.639	0.645	0.653	0.741
		0.752	0.638	0.723	0.694	0.751
		0.706	0.644	0.723	0.597	0.703
	MSP	0.632	0.615	0.576	0.645	0.686
		0.632	0.524	0.596	0.530	0.700
		0.628	0.503	0.608	0.429	0.670
	MI	0.505	0.510	0.537	0.495	0.505
		0.500	0.486	0.498	0.486	0.501
		0.505	0.511	0.502	0.507	0.509
Spectrum-Llama-3.1-8B	all UQ	0.765	0.764	0.764	0.762	0.764
		0.766	0.764	0.766	0.762	0.763
		0.765	0.764	0.764	0.762	0.765
	CoCoA	0.727	0.727	0.734	0.725	0.727
		0.722	0.724	0.733	0.724	0.724
		0.731	0.731	0.734	0.728	0.731
	Semantic energy	0.721	0.722	0.722	0.715	0.721
		0.716	0.716	0.719	0.719	0.718
		0.722	0.723	0.724	0.717	0.724
	Semantic entropy	0.725	0.723	0.725	0.699	0.725
		0.726	0.723	0.726	0.714	0.726
		0.721	0.719	0.722	0.688	0.722
	SAR	0.755	0.755	0.755	0.750	0.755
		0.753	0.753	0.753	0.751	0.753
		0.757	0.757	0.757	0.751	0.757
	MSP	0.701	0.703	0.702	0.692	0.701
		0.696	0.699	0.696	0.693	0.696
		0.706	0.706	0.706	0.686	0.706
	MI	0.578	0.592	0.586	0.552	0.578
		0.569	0.581	0.565	0.546	0.565
		0.589	0.599	0.579	0.563	0.592
Spectrum-Qwen3-14B	all UQ	0.813	0.811	0.815	0.811	0.814
		0.817	0.816	0.819	0.815	0.818
		0.807	0.805	0.808	0.801	0.808
	CoCoA	0.789	0.788	0.789	0.778	0.789
		0.794	0.794	0.795	0.789	0.795
		0.780	0.778	0.781	0.769	0.781
	Semantic energy	0.774	0.773	0.774	0.766	0.774
		0.782	0.782	0.783	0.777	0.783
		0.763	0.761	0.762	0.757	0.762
	Semantic entropy	0.778	0.778	0.778	0.772	0.778
		0.787	0.787	0.787	0.782	0.787
		0.766	0.765	0.766	0.760	0.766
	SAR	0.795	0.795	0.796	0.787	0.795
		0.801	0.802	0.801	0.795	0.801
		0.785	0.784	0.785	0.776	0.785
	MSP	0.763	0.762	0.763	0.757	0.763
		0.767	0.766	0.767	0.763	0.767
		0.757	0.755	0.757	0.750	0.757
	MI	0.584	0.587	0.589	0.589	0.579
		0.567	0.564	0.559	0.553	0.553
		0.606	0.602	0.600	0.602	0.602
A2Search-7B	all UQ	0.738	0.740	0.746	0.738	0.737
		0.735	0.731	0.738	0.732	0.734
		0.746	0.737	0.751	0.737	0.743
	CoCoA	0.735	0.736	0.740	0.700	0.735
		0.728	0.729	0.729	0.694	0.728
		0.746	0.745	0.746	0.710	0.746
	Semantic energy	0.724	0.727	0.728	0.726	0.724
		0.724	0.726	0.724	0.714	0.724
		0.728	0.729	0.728	0.721	0.727
	Semantic entropy	0.727	0.729	0.729	0.724	0.726
		0.725	0.725	0.725	0.722	0.725
		0.729	0.732	0.729	0.729	0.729
	SAR	0.606	0.684	0.644	0.569	0.605
		0.344	0.620	0.520	0.649	0.590
		0.361	0.581	0.525	0.672	0.617
	MSP	0.657	0.547	0.506	0.542	0.657
		0.360	0.374	0.348	0.528	0.351
		0.375	0.389	0.531	0.555	0.348
	MI	0.502	0.543	0.526	0.543	0.502
		0.504	0.551	0.497	0.550	0.495
		0.500	0.526	0.503	0.526	0.509

Table 18: AUROC (error prediction) on AmbigQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best AUROC for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.904	0.904	0.904	0.904	0.903
		0.913	0.914	0.913	0.914	0.914
		0.891	0.892	0.891	0.891	0.891
	CoCoA	0.884	0.884	0.883	0.884	0.884
		0.892	0.893	0.892	0.893	0.892
		0.871	0.872	0.871	0.872	0.871
	Semantic energy	0.874	0.875	0.874	0.876	0.874
		0.885	0.886	0.885	0.887	0.885
		0.858	0.860	0.858	0.860	0.858
	Semantic entropy	0.880	0.881	0.879	0.881	0.880
		0.891	0.892	0.891	0.893	0.891
		0.865	0.866	0.865	0.866	0.865
	SAR	0.892	0.892	0.892	0.893	0.892
		0.902	0.903	0.902	0.903	0.902
		0.878	0.879	0.878	0.879	0.878
	MSP	0.853	0.855	0.853	0.857	0.853
		0.858	0.861	0.858	0.864	0.858
		0.843	0.845	0.843	0.846	0.843
	MI	0.864	0.867	0.864	0.869	0.863
		0.872	0.876	0.872	0.877	0.872
		0.850	0.854	0.850	0.856	0.850
Qwen3-14B	all UQ	0.859	0.863	0.860	0.871	0.857
		0.868	0.871	0.868	0.877	0.867
		0.847	0.851	0.847	0.860	0.845
	CoCoA	0.834	0.841	0.835	0.847	0.834
		0.841	0.846	0.841	0.851	0.841
		0.824	0.831	0.824	0.837	0.824
	Semantic energy	0.840	0.846	0.842	0.854	0.840
		0.852	0.856	0.852	0.861	0.852
		0.826	0.832	0.825	0.842	0.826
	Semantic entropy	0.811	0.840	0.818	0.842	0.810
		0.818	0.844	0.819	0.846	0.819
		0.800	0.830	0.799	0.831	0.799
	SAR	0.842	0.854	0.846	0.858	0.842
		0.850	0.858	0.852	0.860	0.850
		0.832	0.845	0.834	0.849	0.832
	MSP	0.820	0.830	0.822	0.835	0.820
		0.828	0.836	0.828	0.840	0.828
		0.809	0.820	0.809	0.823	0.809
	MI	0.539	0.735	0.586	0.735	0.539
		0.546	0.730	0.546	0.730	0.546
		0.532	0.727	0.532	0.727	0.532
Spectrum-Llama-3.1-8B	all UQ	0.887	0.888	0.887	0.889	0.887
		0.898	0.899	0.897	0.900	0.898
		0.873	0.874	0.873	0.875	0.873
	CoCoA	0.857	0.859	0.857	0.861	0.857
		0.866	0.868	0.866	0.869	0.866
		0.845	0.846	0.845	0.848	0.845
	Semantic energy	0.858	0.861	0.859	0.863	0.858
		0.867	0.870	0.867	0.872	0.867
		0.845	0.849	0.845	0.849	0.845
	Semantic entropy	0.864	0.867	0.864	0.869	0.864
		0.874	0.877	0.874	0.878	0.874
		0.851	0.854	0.851	0.854	0.851
	SAR	0.875	0.876	0.875	0.876	0.875
		0.884	0.885	0.884	0.885	0.884
		0.863	0.864	0.863	0.864	0.863
	MSP	0.811	0.818	0.812	0.825	0.811
		0.816	0.823	0.816	0.833	0.816
		0.802	0.808	0.802	0.812	0.802
	MI	0.843	0.849	0.838	0.850	0.842
		0.849	0.856	0.849	0.856	0.849
		0.832	0.838	0.825	0.836	0.832
Spectrum-Qwen3-14B	all UQ	0.906	0.907	0.906	0.907	0.906
		0.916	0.916	0.915	0.916	0.915
		0.894	0.894	0.894	0.894	0.894
	CoCoA	0.887	0.888	0.887	0.888	0.887
		0.895	0.896	0.895	0.896	0.895
		0.875	0.875	0.875	0.876	0.875
	Semantic energy	0.877	0.879	0.877	0.881	0.877
		0.887	0.888	0.887	0.890	0.887
		0.864	0.865	0.864	0.867	0.864
	Semantic entropy	0.884	0.886	0.885	0.888	0.884
		0.894	0.896	0.894	0.897	0.894
		0.871	0.873	0.871	0.874	0.871
	SAR	0.895	0.896	0.895	0.897	0.895
		0.904	0.904	0.904	0.905	0.904
		0.883	0.884	0.883	0.884	0.883
	MSP	0.852	0.855	0.853	0.857	0.852
		0.857	0.861	0.857	0.864	0.857
		0.843	0.845	0.843	0.846	0.843
	MI	0.871	0.875	0.866	0.875	0.867
		0.873	0.881	0.873	0.882	0.873
		0.856	0.864	0.856	0.863	0.856
A2Search-7B	all UQ	0.838	0.843	0.839	0.850	0.838
		0.847	0.851	0.846	0.856	0.846
		0.828	0.833	0.829	0.841	0.828
	CoCoA	0.808	0.819	0.809	0.824	0.808
		0.816	0.825	0.816	0.830	0.816
		0.797	0.809	0.797	0.812	0.797
	Semantic energy	0.825	0.830	0.826	0.837	0.824
		0.834	0.838	0.834	0.844	0.833
		0.814	0.819	0.814	0.826	0.814
	Semantic entropy	0.804	0.826	0.809	0.828	0.804
		0.810	0.831	0.810	0.832	0.810
		0.796	0.817	0.796	0.819	0.796
	SAR	0.822	0.832	0.823	0.837	0.822
		0.829	0.837	0.829	0.841	0.829
		0.813	0.824	0.813	0.827	0.813
	MSP	0.794	0.801	0.797	0.812	0.794
		0.803	0.810	0.803	0.818	0.803
		0.783	0.792	0.783	0.799	0.783
	MI	0.545	0.696	0.585	0.700	0.545
		0.545	0.692	0.548	0.697	0.547
		0.546	0.690	0.546	0.691	0.545

Table 19: AUROC (error prediction) on TriviaQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best AUROC for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.718	0.846	0.747	0.840	0.741
		1.000	0.824	0.235	0.765	0.824
		0.654	0.846	0.875	0.856	0.734
	CoCoA	0.676	0.684	0.682	0.478	0.558
		1.000	0.647	0.471	0.647	0.000
		0.609	0.679	0.763	0.449	0.705
	Semantic energy	0.665	0.672	0.701	0.653	0.598
		0.176	0.412	1.000	0.118	0.176
		0.772	0.744	0.763	0.776	0.712
	Semantic entropy	0.533	0.716	0.592	0.722	0.541
		0.706	0.824	0.235	0.765	0.235
		0.519	0.696	0.638	0.708	0.647
	SAR	0.653	0.733	0.733	0.621	0.686
		0.118	0.294	0.882	0.235	0.176
		0.788	0.846	0.817	0.712	0.811
	MSP	0.823	0.792	0.739	0.745	0.844
		0.941	0.647	0.882	0.765	0.941
		0.788	0.814	0.827	0.740	0.821
	MI	0.595	0.760	0.639	0.731	0.588
		0.735	0.706	0.324	0.588	0.735
		0.564	0.779	0.705	0.772	0.561
Qwen3-14B	all UQ	0.794	0.927	0.799	0.927	0.823
		0.500	0.500	0.500	0.500	0.500
		0.785	0.929	0.840	0.926	0.821
	CoCoA	0.914	0.661	0.731	0.656	0.917
		0.500	0.500	0.500	0.500	0.500
		0.920	0.683	0.901	0.683	0.929
	Semantic energy	0.923	0.818	0.714	0.708	0.762
		0.500	0.500	0.500	0.500	0.500
		0.923	0.827	0.878	0.702	0.769
	Semantic entropy	0.542	0.521	0.604	0.562	0.542
		0.500	0.500	0.500	0.500	0.500
		0.538	0.535	0.538	0.535	0.538
	SAR	0.823	0.680	0.747	0.661	0.754
		0.500	0.500	0.500	0.500	0.500
		0.840	0.683	0.715	0.660	0.742
	MSP	0.905	0.927	0.879	0.716	0.900
		0.500	0.500	0.500	0.500	0.500
		0.904	0.926	0.877	0.708	0.902
	MI	0.510	0.661	0.422	0.641	0.510
		0.500	0.500	0.500	0.500	0.500
		0.506	0.673	0.506	0.660	0.506
Spectrum-Llama-3.1-8B	all UQ	0.605	0.575	0.578	0.739	0.605
		0.412	0.118	0.235	1.000	0.353
		0.636	0.642	0.636	0.683	0.657
	CoCoA	0.592	0.631	0.523	0.580	0.569
		0.294	0.882	0.353	0.941	0.294
		0.652	0.584	0.551	0.538	0.639
	Semantic energy	0.653	0.683	0.626	0.647	0.639
		1.000	0.647	0.588	0.824	0.588
		0.577	0.694	0.631	0.636	0.652
	Semantic entropy	0.470	0.594	0.704	0.567	0.514
		0.118	0.824	0.647	0.824	0.471
		0.558	0.545	0.764	0.540	0.522
	SAR	0.837	0.816	0.723	0.761	0.837
		1.000	1.000	0.529	1.000	1.000
		0.813	0.790	0.738	0.719	0.813
	MSP	0.534	0.583	0.608	0.553	0.527
		0.059	0.941	0.941	0.706	0.059
		0.631	0.522	0.649	0.530	0.623
	MI	0.732	0.668	0.565	0.695	0.713
		0.235	0.059	0.294	0.235	0.000
		0.832	0.745	0.609	0.787	0.858
Spectrum-Qwen3-14B	all UQ	0.857	0.905	0.897	0.909	0.848
		0.588	1.000	0.588	0.941	0.412
		0.933	0.891	0.939	0.933	0.936
	CoCoA	0.676	0.800	0.674	0.705	0.701
		0.824	1.000	1.000	1.000	0.824
		0.667	0.756	0.580	0.644	0.699
	Semantic energy	0.684	0.779	0.796	0.756	0.705
		0.353	0.647	0.471	0.765	0.412
		0.750	0.814	0.865	0.760	0.760
	Semantic entropy	0.716	0.728	0.657	0.640	0.720
		0.471	0.529	0.294	0.000	0.176
		0.760	0.772	0.734	0.760	0.830
	SAR	0.691	0.686	0.722	0.661	0.710
		0.588	0.588	0.706	0.176	0.529
		0.718	0.715	0.737	0.769	0.753
	MSP	0.756	0.794	0.724	0.760	0.739
		0.706	1.000	0.647	1.000	0.706
		0.782	0.747	0.817	0.696	0.756
	MI	0.695	0.750	0.754	0.699	0.720
		0.324	0.529	0.324	0.529	0.324
		0.806	0.814	0.822	0.750	0.822
A ² Search-7B	all UQ	0.739	0.737	0.587	0.699	0.844
		0.500	0.500	0.500	0.500	0.500
		0.756	0.719	0.686	0.704	0.834
	CoCoA	0.637	0.659	0.480	0.695	0.673
		0.500	0.500	0.500	0.500	0.500
		0.662	0.652	0.591	0.686	0.681
	Semantic energy	0.731	0.646	0.658	0.638	0.710
		0.500	0.500	0.500	0.500	0.500
		0.729	0.660	0.812	0.644	0.716
	Semantic entropy	0.574	0.617	0.453	0.530	0.574
		0.500	0.500	0.500	0.500	0.500
		0.571	0.600	0.571	0.543	0.571
	SAR	0.848	0.674	0.676	0.545	0.817
		0.500	0.500	0.500	0.500	0.500
		0.864	0.678	0.808	0.538	0.829
	MSP	0.740	0.710	0.539	0.682	0.687
		0.500	0.500	0.500	0.500	0.500
		0.726	0.709	0.665	0.675	0.708
	MI	0.526	0.577	0.432	0.575	0.516
		0.500	0.500	0.500	0.500	0.500
		0.532	0.579	0.532	0.577	0.532

Table 20: AUROC (error prediction) on NCQA (test split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, paired bootstrap). **Highlight** = best AUROC for model.