The Role of Ambiguity in Error Prediction via Uncertainty Quantification

Anonymous ACL submission

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, which is inherent in the model input and context. This paper presents a method for improving error prediction for Large Language Models (LLMs), by disentangling input ambiguity from UQ signal. We conduct experiments on the task of 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 scores across model families, training and evaluation paradigms, datasets (including allegedly unambiguous ones), and sources of aleatoric uncertainty, yielding improvements of over 10 points of PRR for individual UQ metrics on standard datasets.

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).

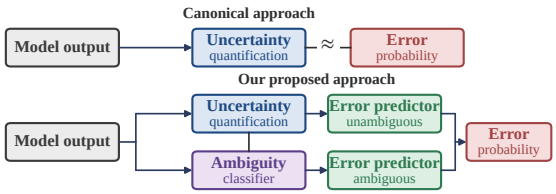

Figure 1: The established error prediction method directly using model features such as uncertainty metrics (top) and our proposed framework, predicting ambiguity and using it in a gated experts model for improving error prediction scores (bottom).

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 incorrectly predict high probability of error due to high model uncertainty, as it does not discriminate aleatoric and epistemic uncertainty.

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 uncertainty reflecting a model's lack of knowledge, and aleatoric uncertainty reflecting inherent input am-

biguity 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 is useful in predicting errors on both ambiguous and unambiguous QA datasets, for various model families, including models that have been specifically finetuned to represent human label variation, for all uncertainty metrics considered, including their combinations.[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 Language 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.

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 uncertainty 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

[1]The code will be released on https://anonymous.4open.science/r/error_or_ambiguity-F682/

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. Ahdritz 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

This section describes the approach proposed in this paper, namely extending the standard approach of using UQ metrics directly as error predictors (which we define formally as our baseline in Section 4.4, and to which we compare 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 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. That is, UQ scores track error rates strongly on unambiguous inputs but only weakly on ambiguous ones, where error rates vary less with uncertainty. Conditioning on ambiguity lets the model trust UQ where it is informative and discount it where it is not.

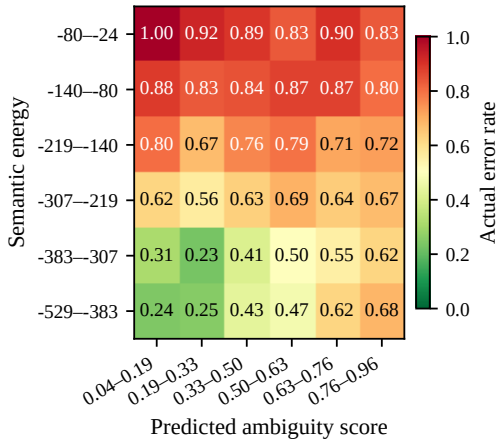

Figure 2: Heatmap illustrating the relationship between UQ (Semantic Entropy) and predicted ambiguity on the one hand, and the error rate on the other. Ambiguity scores are predicted by the model described in Section 3.2; the same pattern holds with gold ambiguity labels, confirming it reflects a genuine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

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 samples 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 unnormalised 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.

To assess which features are most informative for ambiguity classification, we measure the AUROC of each scalar feature as a standalone binary classifier of the ambiguity label (ambiguous vs. unambiguous), 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. Representation features show stronger correlations than scalar features, particularly on AmbigQA. These results motivate their use as features in the ambiguity classifier described in Section 3.2.

3.1.3 Internal Representations

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

3.2 The Role of Ambiguity in Error Prediction via UQ

Having established the features used for ambiguity classification and error prediction, we now describe the model architectures that disentangle the ambiguity signal in error prediction. Both build on the baseline error prediction head (Section 4.4), which

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

	Internal		UQ						Semantic Cluster					
Data	Emb	Res	SE	SEng	MSP	SAR	MI	CoCoA	#Cl	ClE	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 = layer-15 residual. #Cl = number of clusters; ClE = cluster entropy; MxC = max cluster probability; ENA = effective number of answers; PG2 = probability gap (top-2); Top2 = top-2 probability mass.

maps a UQ feature vector $\mathbf{x}_f$ to a scalar error logit $\hat{c}$; the architectures below differ only in how they incorporate the ambiguity signal.

The **Gated Experts** model routes instances to specialized expert networks for ambiguous (E_a) and unambiguous (E_u) queries based on ambiguity signal to mitigate feature interference (see bottom of Figure 1 for illustration). The **Oracle Version** hard-routes using ground-truth ambiguity labels $a \in \{0, 1\}$ such that $\hat{c} = (1 - a)\, E_u(\mathbf{x}_f) + a\, E_a(\mathbf{x}_f)$. The **Latent Version** soft-routes using the pre-trained ambiguity model probability $\hat{a}$ such that $\hat{c} = (1 - \hat{a})\, E_u(\mathbf{x}_f) + \hat{a}\, E_a(\mathbf{x}_f)$.

The **Selective Prediction Framework** (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017) computes a rejection score R that penalizes the error logit $\hat{c}$ proportionally to ambiguity, scaled by λ, used to rank instances for abstention. Unlike Cole et al. (2023), who use sampling-based confidence scores to threshold abstention, our selective prediction framework explicitly incorporates ambiguity as a penalty on the error score, allowing the model to abstain preferentially on instances that are both uncertain and ambiguous. The **Oracle Version** uses ground-truth ambiguity a yielding $R = -\hat{c} + \lambda\, a$, whereas the **Latent Version** uses the pre-trained ambiguity model probability $\hat{a}$ yielding $R = -\hat{c} + \lambda\, \hat{a}$.

The Ambiguity Model is trained independently of the error prediction pipeline. It takes the same scalar UQ features $\mathbf{x}_f$ as input, processes them through an MLP to obtain a latent representation $\mathbf{z}_a = \text{MLP}_a(\mathbf{x}_f)$, and predicts the probability of instance-level semantic ambiguity via a linear classifier $\hat{a} = \sigma(\mathbf{w}^\top \mathbf{z}_a + b)$, optimized using binary cross-entropy against ground-truth ambiguity labels. The resulting ambiguity probabilities $\hat{a}$ are used as gating or penalty signals in the models described above.

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

This section describes how we tested the contributions of ambiguity signal in predicting errors. We discuss the experimental details of running LLMs for inference on QA datasets, evaluated their QA performance and measured UQ scores, which features we then used to predict errors and how we evaluated the latter task.

4.1 Data

The **AmbigQA** dataset (Min et al., 2020) formalizes ambiguity in open-domain QA by extending NQ (Kwiatkowski et al., 2019) with multiple plausible answer sets. The authors find that over 50% of natural queries contain inherent ambiguity, such as 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 (Bai et al., 2023) model with Thinking Mode. Huang et al. (2026) validate the quality of this annotation approach against human judgements.

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 representing three distinct paradigms: 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** (Bai et al., 2023). 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 maintain the base architectures but undergo Spectrum tuning to mitigate mode collapse. This enables them to preserve distributional coverage and generate diverse, valid responses to ambiguous prompts.

Finally, we benchmark against **A^{2}Search-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 (Bai et al., 2023).

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 (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 fol-

low 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 metrics $\mathbf{x}_f$, 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 is optimized with a composite objective $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \alpha\mathcal{L}_{\text{rank}}$, where $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy (BCE) loss and $\mathcal{L}_{\text{rank}} = \frac{1}{|P||N|}\sum_{i\in P, j\in N}\log(1+\exp(\hat{c}_j - \hat{c}_i))$ 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 quantifies the effectiveness of using uncertainty scores to abstain from answering. PRR normalizes the area under the accuracy-rejection curve by comparing the model's rejection strategy against both a random baseline and an optimal oracle, offering a standardized metric for selective generation quality.

Statistical significance of improvements over the baseline is assessed using a permutation test ($p < 0.05$).

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 median-layer hidden state 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 30 epochs, using early stopping when PRR does not improve, with patience of 5 epochs. We carried out a hyperparameter search with hidden dimensions in $\{128, 256, 512, 1024, 2048\}$ and learning rates in $\{3\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}\}$. We report the best score from the search for each model configuration. Following standard practice in the UQ literature (Vashurin et al., 2025; Farquhar et al., 2024), we report results on the validation splits of AmbigQA and TriviaQA, as the test set labels for both datasets are withheld. NCQA results are reported on its held-out test set.

5 Results

We present our main findings of the role of ambiguity in error prediction via UQ in Section 5.3. To lay the ground for them, in Section 5.1 we demonstrate the model performance on the QA task, while Section 5.2 presents the performance of ambiguity prediction models.

5.1 QA Performance

QA output evaluation reveals consistent differences between ambiguous and unambiguous instances across all models and datasets. 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, we note that all models score much higher on both evaluation metrics for unambiguous instances. This result indicates that ambiguous ones are harder in general, because the evaluation metrics use a maximum score between all plausible reference answers, which gives an advantage to ambiguous questions. Notably, 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.

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^2Search 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^2Search-7B	0.693	0.623	0.642

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

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 extract robust patterns.

5.3 Error Prediction

The difference in UQ metric ability to predict model errors is presented in Table 5.[3] We find that ambiguity signal improves error prediction across models, datasets, UQ metrics, evaluation metrics, feature sets and architectures of error prediction models. Notably, the latent models often match or exceed their oracle counterparts despite using predicted rather than gold ambiguity labels. This suggests that the continuous ambiguity probabilities produced by the ambiguity classifier carry a more useful signal than the binary gold labels.

Figure 3 illustrates the improvements introduced by our proposed models. Each curve traces what happens as we accept progressively more of the model's answers, ordered from most to least confident. The y-axis shows the error rate among accepted answers, so a well-calibrated uncertainty measure should keep errors low until coverage is high. Curves are plotted separately for ambiguous and unambiguous questions, with the upper-bound lines showing what would be achievable if the uncertainty measure ranked every incorrect answer last. First, the improvement appears across both ambiguous and unambiguous instances, indicating that the proposed architecture meaningfully encodes the two types of relationships, namely the one between epistemic uncertainty and error likelihood (unambiguous), and the one between a combination of aleatoric and epistemic uncertainty (ambiguous) and error likelihood. 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.

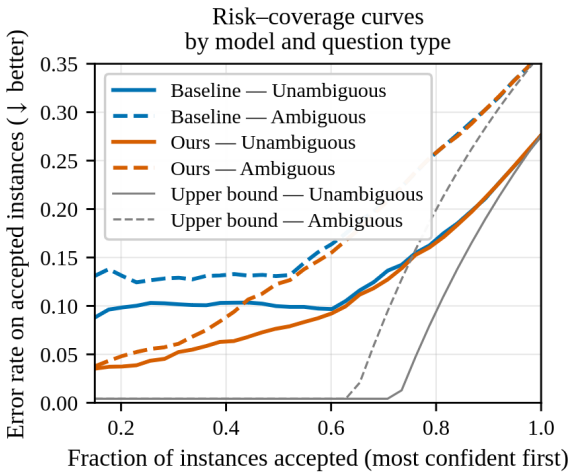

Figure 3: Risk–coverage curves for baseline and latent selective models on unambiguous and ambiguous questions, with upper bound (Qwen3-14B, TriviaQA).

Nonetheless, interesting differences emerge between the UQ metrics themselves. First, we find that Maximum Softmax Probability (MSP) is surprisingly robust. While complex semantic clustering methods are generally assumed to be superior, MSP frequently outperforms Semantic Entropy on inherently ambiguous datasets. This reinforces the premise that semantic clustering metrics degrade when valid aleatoric variations exist in the data. Among the individual UQ features, CoCoA and

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

SAR are in the lead. Notably, they still benefit from explicit ambiguity routing. 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). Crucially, we observe the highest overall error prediction performance when combining all UQ features into a single representation (*all UQ*). Yet, even when the error prediction model has access to an ensemble of every UQ metric, applying our ambiguity signal via latent gating or selective prediction still yields consistent PRR improvements. This supports our core hypothesis that UQ metrics (even in combination) do not distinguish aleatoric and epistemic uncertainty, and our ambiguity classifier provides a complementary signal for establishing model reliability.

Comparing the different datasets, the NCQA improvements are the most dramatic due to the fact that, in contrast to other datasets, the ambiguous cases are the easy ones to solve given that both answers to a binary question would be deemed correct. We note that the improvements on NCQA should be interpreted with caution given the small training set of 234 instances. However, the results are reported on the held-out test split, and the pattern is consistent across feature sets and model families, suggesting it is not an artefact of overfitting to a single configuration.

Finally, models finetuned to represent multiple plausible answers (e.g. Spectrum-Qwen) are only marginally better at error prediction than their vanilla counterparts, suggesting such finetuning improves calibration alongside steerability but leaves room for further gains. These models still benefit from latent gating, indicating that ambiguity-aware finetuning and explicit ambiguity signal are not redundant.

To illustrate, consider the pair of questions in Table 6. 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 73.8% predicted error to the ambiguous question. Our latent selective model disentangles underspecification from error, assigning near-zero predicted error (0.0%) while giving a similarly low prediction 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.

Dataset	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
AmbigQA	all UQ	0.555	**0.571**	**0.575**	**0.562**	**0.537**
	CoCoA	0.538	**0.553**	**0.553**	0.524	0.526
	Sem. energy	0.392	**0.520**	**0.460**	**0.520**	**0.432**
	Sem. entropy	0.421	**0.513**	**0.471**	**0.508**	**0.446**
	SAR	0.525	**0.539**	**0.543**	0.526	0.528
	MSP	0.483	**0.512**	**0.504**	**0.492**	0.478
	MI	0.014	**0.078**	**0.058**	**0.078**	**0.058**
TriviaQA	all UQ	0.787	**0.795**	**0.788**	**0.813**	**0.778**
	CoCoA	0.753	**0.773**	0.752	**0.784**	**0.742**
	Sem. energy	0.754	**0.778**	**0.757**	**0.793**	0.748
	Sem. entropy	0.641	**0.759**	**0.673**	**0.761**	**0.671**
	SAR	0.735	**0.773**	**0.739**	**0.786**	0.729
	MSP	0.738	**0.758**	**0.738**	**0.771**	**0.725**
	MI	0.204	**0.568**	0.196	**0.569**	0.197
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 5: PRR on validation split (test for NCQA) for Qwen3-14B (AlignScore supervision). **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best PRR for model/dataset.

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 6: Predicted error probability. Acc = alignscore of the most likely response. Base = semantic entropy only. Ours = latent selective. $\hat{p}_{\text{amb}}$ = predicted ambiguity.

6 Conclusion

In this paper we claim that ambiguity signal is useful for predicting errors from uncertainty, due to the fact that uncertainty encompasses both epistemic and aleatoric sources. Experimental results show that ambiguity information is beneficial regardless of the model, UQ metric, dataset, feature set and architecture choice for error prediction. We show that using UQ metrics along with ambiguity can yield high error prediction performance, up to 0.879 PRR and 0.904 AUROC on TriviaQA. These results contribute to improving the reliability of LLMs. Future work could extend this framework to retrieval-augmented settings, where retrieved context introduces a further source of aleatoric uncertainty, or to settings where ambiguity labels must be induced fully unsupervised.

Limitations

The study only covers textual data, and all questions and answers are in the English language, hence further work would need to establish whether the discovered 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. Extending our approach to settings where ambiguity labels must be fully automatically induced remains an important direction for future work.

Furthermore, the study focuses on semantic uncertainty quantification metrics, which is one of many types of measures of uncertainty. While the semantically rich metrics achieve state-of-the-art results and are therefore the most interesting to study, other work could explore whether other UQ metrics would have a similar relationship to ambiguity.

In addition, hyperparameter selection is performed by evaluating PRR on the same validation set used for final reporting. While this is a common practice in error prediction research due to the absence of held-out test splits for AmbigQA and TriviaQA (whose test labels are withheld for leaderboard evaluation), it may lead to optimistic reported scores; results should be interpreted accordingly.

Finally, this research paper is limited to closed-book QA, which restricts the knowledge available to the model to that which was present in the training data. It would be of great interest to extend the work to a retrieval-augmented-generation setup and compare the uncertainty and ambiguity relationship to a setup where more information is available via the retrieval process, especially the cases where the retrieved data contradicts internal knowledge.

Acknowledgments

References

Gustaf Ahdritz, 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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

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.

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 Cattan, 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 (UncertaiNLP 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.

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 Mireshghallah, 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.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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.

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. A^2 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 7 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.

		Internal Repr.		UQ					Sem. Clust.						
Model	**Dataset**	**Emb**	**Res**	**SE**	**SEng**	**MSP**	**SAR**	**MI**	**CoCoA**	**#Cl**	**ClEnt**	**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^2Search 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 7: Same as Table 1 for Spectrum-Llama, Qwen3-14B, Spectrum-Qwen3-14B and A^2Search-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 8–19 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 (gated or selective, latent or oracle) improve over the UQ-only baseline for the large majority of model–feature combinations, with the latent variants typically tracking or exceeding the oracle. The few non-significant or negative entries are concentrated on NCQA, where the small training set yields high variance, and on MI, whose baseline performance is already close to chance.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.632 0.722 0.504	0.633 0.722 0.502	**0.651** 0.727 0.511	0.620 0.708 0.483	0.627 0.721 0.502
	CoCoA	0.615 0.706 0.490	0.613 0.703 0.487	**0.629** 0.706 0.490	0.611 0.700 0.481	0.614 0.706 0.490
	Semantic energy	0.593 0.684 0.462	0.596 0.682 0.463	**0.606** 0.682 0.461	0.585 0.668 0.450	0.585 0.682 0.461
	Semantic entropy	0.595 0.677 0.463	0.595 0.680 0.463	**0.611** 0.680 0.463	0.583 0.665 0.446	0.589 0.680 0.462
	SAR	0.590 0.685 0.461	0.593 0.684 0.447	**0.608** 0.685 0.459	0.589 0.678 0.435	**0.598** 0.685 0.459
	MSP	0.552 0.643 0.432	0.553 0.639 0.433	**0.567** 0.642 0.430	0.539 0.622 0.412	0.550 0.642 0.430
	MI	0.189 0.164 0.156	0.189 0.184 0.167	**0.199** 0.159 0.144	0.181 0.178 0.156	0.182 0.155 0.158
Qwen3-14B	all UQ	0.555 0.645 0.433	**0.571** 0.663 0.428	**0.575** 0.658 0.429	**0.562** 0.650 0.417	0.537 0.642 0.423
	CoCoA	0.538 0.651 0.380	**0.553** 0.657 0.392	**0.553** 0.649 0.379	0.524 0.611 0.379	0.526 0.649 0.379
	Semantic energy	0.392 0.498 0.261	**0.520** 0.615 0.357	**0.460** 0.518 0.286	**0.520** 0.613 0.357	**0.432** 0.498 0.261
	Semantic entropy	0.421 0.538 0.296	**0.513** 0.594 0.374	**0.471** 0.538 0.295	**0.508** 0.589 0.371	**0.446** 0.538 0.295
	SAR	0.525 0.625 0.386	0.539 0.632 0.387	**0.543** 0.624 0.386	0.526 0.604 0.378	0.528 0.624 0.386
	MSP	0.483 0.599 0.324	**0.512** 0.618 0.344	**0.504** 0.597 0.323	**0.492** 0.586 0.338	0.478 0.597 0.323
	MI	0.014 -0.026 0.019	**0.078** 0.057 0.083	**0.058** -0.026 0.014	**0.078** 0.057 0.083	**0.058** -0.026 0.020
Spectrum-Llama-3.1-8B	all UQ	0.461 0.515 0.394	0.462 0.515 0.385	**0.483** 0.535 0.401	0.457 0.510 0.382	0.450 0.516 0.399
	CoCoA	0.465 0.504 0.403	0.463 0.503 0.401	0.466 0.504 0.403	0.448 0.492 0.375	0.429 0.504 0.403
	Semantic energy	0.455 0.509 0.371	0.456 0.509 0.368	0.460 0.509 0.370	0.404 0.452 0.319	0.411 0.509 0.370
	Semantic entropy	0.459 0.517 0.371	0.457 0.509 0.366	0.468 0.517 0.371	0.410 0.466 0.327	0.396 0.515 0.371
	SAR	0.461 0.509 0.386	0.462 0.508 0.383	0.464 0.508 0.386	0.457 0.502 0.373	0.442 0.508 0.386
	MSP	0.417 0.451 0.363	0.417 0.455 0.356	0.417 0.450 0.365	0.404 0.445 0.340	0.379 0.450 0.365
	MI	0.239 0.198 0.132	0.218 0.228 0.197	0.232 0.192 0.134	0.216 0.227 0.194	0.215 0.196 0.141
Spectrum-Qwen3-14B	all UQ	0.560 0.603 0.505	0.560 0.600 0.505	0.565 0.602 0.511	0.556 0.599 0.496	0.536 0.602 0.507
	CoCoA	0.538 0.567 0.493	0.538 0.567 0.495	0.537 0.567 0.493	0.518 0.556 0.459	0.492 0.567 0.493
	Semantic energy	0.507 0.557 0.431	**0.526** 0.570 0.450	**0.520** 0.558 0.431	**0.521** 0.567 0.445	0.487 0.557 0.430
	Semantic entropy	0.545 0.583 0.479	0.542 0.581 0.479	**0.552** 0.583 0.479	0.536 0.577 0.472	0.508 0.581 0.479
	SAR	0.510 0.565 0.429	**0.525** 0.576 0.436	0.516 0.565 0.429	**0.523** 0.573 0.434	0.493 0.565 0.429
	MSP	0.494 0.520 0.454	0.497 0.526 0.454	0.491 0.516 0.452	0.483 0.515 0.434	0.451 0.517 0.452
	MI	0.265 0.194 0.159	0.245 0.258 0.219	0.247 0.193 0.139	0.245 0.260 0.217	0.245 0.191 0.154
A^2Search-7B	all UQ	0.543 0.640 0.409	**0.580** 0.668 0.429	**0.573** 0.655 0.419	**0.580** 0.664 0.441	0.545 0.641 0.401
	CoCoA	0.498 0.607 0.369	**0.558** 0.629 0.433	**0.537** 0.606 0.367	**0.550** 0.610 0.438	**0.512** 0.606 0.367
	Semantic energy	0.405 0.528 0.291	**0.556** 0.643 0.405	**0.475** 0.534 0.283	**0.557** 0.645 0.408	**0.463** 0.527 0.290
	Semantic entropy	0.488 0.580 0.379	**0.555** 0.630 0.414	**0.534** 0.560 0.344	**0.557** 0.632 0.418	**0.510** 0.561 0.345
	SAR	0.489 0.588 0.366	**0.544** 0.614 0.419	**0.511** 0.588 0.367	**0.544** 0.614 0.421	**0.497** 0.588 0.367
	MSP	0.447 0.568 0.320	**0.539** 0.625 0.398	**0.495** 0.562 0.318	**0.542** 0.625 0.404	**0.477** 0.562 0.318
	MI	0.025 -0.003 0.040	**0.147** 0.141 0.129	0.056 0.012 0.040	**0.146** 0.140 0.128	0.048 -0.006 0.005

Table 8: PRR on AmbigQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.848 0.858 0.833	0.849 0.859 0.833	0.848 0.856 0.834	0.850 0.859 0.835	0.839 0.857 0.832
	CoCoA	0.829 0.838 0.814	0.831 0.838 0.816	0.830 0.838 0.814	0.831 0.838 0.817	0.817 0.838 0.814
	Semantic energy	0.814 0.822 0.799	0.818 0.824 0.804	0.815 0.822 0.799	0.819 0.825 0.805	0.802 0.822 0.799
	Semantic entropy	0.813 0.815 0.801	**0.820** 0.828 0.806	0.815 0.815 0.800	**0.822** 0.829 0.808	0.804 0.815 0.800
	SAR	0.824 0.832 0.809	0.826 0.832 0.811	0.824 0.832 0.809	0.825 0.830 0.811	0.811 0.832 0.809
	MSP	0.793 0.805 0.774	0.797 0.808 0.779	0.794 0.804 0.774	**0.801** 0.812 0.783	0.785 0.804 0.774
	MI	0.747 0.751 0.743	**0.756** 0.745 0.757	**0.756** 0.751 0.729	**0.763** 0.751 0.764	0.743 0.751 0.729
Qwen3-14B	all UQ	0.787 0.793 0.775	**0.795** 0.799 0.785	0.788 0.793 0.773	**0.813** 0.812 0.805	0.778 0.792 0.775
	CoCoA	0.753 0.757 0.742	**0.773** 0.779 0.760	0.752 0.757 0.742	**0.784** 0.789 0.770	0.742 0.757 0.742
	Semantic energy	0.754 0.763 0.737	**0.778** 0.781 0.767	0.757 0.765 0.736	**0.793** 0.793 0.785	0.748 0.762 0.733
	Semantic entropy	0.641 0.644 0.639	**0.759** 0.758 0.751	**0.673** 0.646 0.637	**0.761** 0.760 0.753	**0.671** 0.644 0.637
	SAR	0.735 0.736 0.726	**0.773** 0.770 0.767	0.739 0.737 0.727	**0.786** 0.781 0.781	0.729 0.736 0.725
	MSP	0.738 0.742 0.728	**0.758** 0.761 0.747	0.738 0.741 0.728	**0.771** 0.775 0.758	0.725 0.741 0.728
	MI	0.204 -0.024 0.004	**0.568** 0.561 0.553	0.196 -0.024 0.004	**0.569** 0.562 0.553	0.197 -0.024 0.004
Spectrum-Llama-3.1-8B	all UQ	0.831 0.845 0.810	0.832 0.846 0.811	0.831 0.845 0.809	0.836 0.850 0.815	0.823 0.845 0.810
	CoCoA	0.804 0.817 0.784	0.808 0.819 0.788	0.805 0.817 0.784	**0.811** 0.822 0.791	0.796 0.817 0.784
	Semantic energy	0.800 0.809 0.784	**0.808** 0.816 0.793	0.801 0.809 0.784	**0.812** 0.822 0.796	0.793 0.809 0.784
	Semantic entropy	0.799 0.811 0.779	**0.812** 0.821 0.796	**0.804** 0.810 0.779	**0.816** 0.825 0.799	0.794 0.810 0.779
	SAR	0.805 0.818 0.784	0.808 0.819 0.789	0.805 0.818 0.784	**0.811** 0.823 0.791	0.795 0.818 0.784
	MSP	0.745 0.762 0.723	**0.756** 0.773 0.733	0.748 0.762 0.722	**0.769** 0.785 0.744	0.739 0.762 0.722
	MI	0.722 0.720 0.707	**0.746** 0.740 0.740	0.717 0.719 0.693	**0.752** 0.748 0.743	0.701 0.719 0.707
Spectrum-Qwen3-14B	all UQ	0.854 0.857 0.844	0.855 0.859 0.845	0.855 0.857 0.844	0.858 0.864 0.847	0.847 0.856 0.844
	CoCoA	0.836 0.841 0.825	0.840 0.846 0.827	0.837 0.840 0.825	0.841 0.848 0.828	0.827 0.840 0.825
	Semantic energy	0.828 0.834 0.815	0.831 0.837 0.818	0.829 0.834 0.816	0.834 0.839 0.821	0.820 0.834 0.815
	Semantic entropy	0.829 0.832 0.814	0.831 0.837 0.820	0.832 0.833 0.820	0.834 0.840 0.821	0.823 0.833 0.820
	SAR	0.840 0.841 0.833	0.843 0.845 0.835	0.841 0.841 0.833	0.844 0.846 0.834	0.830 0.841 0.833
	MSP	0.788 0.794 0.775	0.795 0.803 0.780	0.791 0.794 0.774	**0.803** 0.813 0.785	0.782 0.794 0.774
	MI	0.770 0.765 0.772	**0.785** 0.774 0.787	0.767 0.766 0.771	**0.786** 0.776 0.786	0.753 0.765 0.771
A^2Search-7B	all UQ	0.742 0.767 0.710	**0.755** 0.781 0.721	0.746 0.768 0.711	**0.773** 0.797 0.739	0.737 0.767 0.711
	CoCoA	0.692 0.720 0.656	**0.737** 0.761 0.701	0.698 0.720 0.656	**0.740** 0.766 0.701	0.687 0.720 0.656
	Semantic energy	0.727 0.760 0.687	**0.756** 0.784 0.717	**0.738** 0.763 0.695	**0.767** 0.795 0.726	0.728 0.760 0.687
	Semantic entropy	0.660 0.686 0.625	**0.746** 0.770 0.711	**0.667** 0.682 0.625	**0.747** 0.771 0.711	0.663 0.682 0.625
	SAR	0.704 0.727 0.674	**0.746** 0.770 0.712	0.707 0.727 0.674	**0.754** 0.778 0.718	0.696 0.727 0.674
	MSP	0.675 0.707 0.636	**0.714** 0.741 0.675	**0.684** 0.707 0.636	**0.729** 0.755 0.689	0.674 0.707 0.636
	MI	0.214 -0.002 0.049	**0.480** 0.494 0.444	0.163 -0.002 0.049	**0.482** 0.498 0.443	0.165 -0.002 0.049

Table 9: PRR on TriviaQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). 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^2Search-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 10: PRR on NCQA (test split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.806 0.846 0.754	0.806 0.845 0.753	0.808 0.845 0.755	0.805 0.838 0.745	0.806 0.846 0.753
	CoCoA	0.792 0.835 0.737	0.791 0.834 0.738	0.795 0.836 0.737	0.778 0.825 0.726	0.792 0.835 0.737
	Semantic energy	0.777 0.817 0.727	**0.779** 0.817 0.727	**0.779** 0.816 0.726	0.777 0.817 0.727	0.776 0.816 0.726
	Semantic entropy	0.780 0.818 0.732	0.781 0.819 0.731	**0.783** 0.818 0.732	0.781 0.819 0.730	0.780 0.818 0.732
	SAR	0.784 0.826 0.731	0.785 0.827 0.730	**0.787** 0.826 0.730	0.784 0.826 0.727	0.784 0.826 0.730
	MSP	0.757 0.799 0.705	0.757 0.798 0.705	**0.761** 0.798 0.704	0.757 0.798 0.704	0.756 0.798 0.704
	MI	0.575 0.586 0.569	**0.594** 0.591 0.580	**0.596** 0.586 0.571	0.564 0.572 0.559	0.578 0.584 0.570
Qwen3-14B	all UQ	0.780 0.825 0.719	0.783 0.829 0.717	**0.787** 0.830 0.722	0.782 0.828 0.718	0.776 0.823 0.713
	CoCoA	0.760 0.813 0.694	0.760 0.812 0.693	**0.764** 0.813 0.693	0.763 0.813 0.694	0.760 0.813 0.693
	Semantic energy	0.728 0.778 0.663	**0.752** 0.803 0.679	**0.742** 0.782 0.666	**0.751** 0.802 0.674	0.728 0.778 0.663
	Semantic entropy	0.741 0.789 0.680	**0.755** 0.803 0.687	**0.750** 0.788 0.680	**0.753** 0.803 0.681	0.740 0.788 0.680
	SAR	0.768 0.817 0.702	0.770 0.818 0.702	**0.772** 0.817 0.702	0.772 0.819 0.702	0.768 0.817 0.702
	MSP	0.731 0.783 0.668	**0.738** 0.791 0.667	**0.736** 0.783 0.668	0.704 0.758 0.644	0.731 0.783 0.668
	MI	0.503 0.503 0.502	**0.541** 0.535 0.538	**0.520** 0.503 0.503	**0.541** 0.535 0.538	0.503 0.503 0.502
Spectrum-Llama-3.1-8B	all UQ	0.743 0.769 0.709	0.742 0.768 0.706	**0.748** 0.774 0.713	0.743 0.770 0.706	**0.745** 0.771 0.712
	CoCoA	0.736 0.757 0.708	0.735 0.756 0.707	0.736 0.757 0.708	0.735 0.757 0.703	0.736 0.757 0.708
	Semantic energy	0.728 0.754 0.695	0.730 0.756 0.693	0.730 0.754 0.695	0.693 0.728 0.638	0.728 0.754 0.695
	Semantic entropy	0.730 0.756 0.695	0.730 0.756 0.696	0.731 0.756 0.695	**0.732** 0.758 0.695	0.730 0.756 0.695
	SAR	0.742 0.766 0.708	0.742 0.766 0.708	0.742 0.765 0.708	0.740 0.765 0.705	0.741 0.765 0.708
	MSP	0.706 0.725 0.682	0.706 0.725 0.681	0.706 0.724 0.682	0.706 0.725 0.679	0.706 0.724 0.682
	MI	0.573 0.583 0.572	**0.594** 0.596 0.589	0.579 0.572 0.571	**0.594** 0.595 0.588	0.571 0.572 0.571
Spectrum-Qwen3-14B	all UQ	0.795 0.815 0.770	0.794 0.813 0.769	0.795 0.814 0.770	0.792 0.812 0.766	0.794 0.813 0.769
	CoCoA	0.777 0.794 0.753	0.776 0.792 0.753	0.777 0.794 0.753	0.761 0.779 0.742	0.777 0.794 0.753
	Semantic energy	0.771 0.792 0.743	**0.773** 0.794 0.745	0.772 0.792 0.743	0.769 0.790 0.739	0.770 0.791 0.743
	Semantic entropy	0.780 0.801 0.753	0.781 0.801 0.754	0.781 0.801 0.753	0.780 0.800 0.753	0.780 0.801 0.753
	SAR	0.770 0.795 0.736	**0.776** 0.800 0.741	0.771 0.795 0.736	0.768 0.792 0.734	0.770 0.795 0.736
	MSP	0.750 0.766 0.728	**0.752** 0.769 0.728	0.748 0.763 0.727	0.746 0.761 0.726	0.748 0.763 0.726
	MI	0.589 0.577 0.585	**0.607** 0.614 0.597	0.587 0.576 0.577	**0.605** 0.611 0.596	0.580 0.576 0.585
A^2Search-7B	all UQ	0.785 0.826 0.732	**0.795** 0.833 0.739	0.789 0.827 0.734	**0.792** 0.831 0.738	0.784 0.826 0.729
	CoCoA	0.773 0.808 0.730	**0.785** 0.818 0.739	**0.777** 0.808 0.730	**0.783** 0.815 0.737	0.773 0.808 0.730
	Semantic energy	0.752 0.796 0.701	**0.778** 0.818 0.721	**0.762** 0.796 0.701	**0.777** 0.816 0.718	0.752 0.796 0.701
	Semantic entropy	0.765 0.807 0.713	**0.780** 0.818 0.724	**0.771** 0.806 0.711	**0.780** 0.818 0.724	0.764 0.806 0.712
	SAR	0.760 0.798 0.714	**0.781** 0.814 0.734	**0.763** 0.797 0.714	**0.781** 0.814 0.733	0.760 0.797 0.714
	MSP	0.754 0.795 0.705	**0.770** 0.809 0.716	**0.760** 0.795 0.705	**0.772** 0.811 0.717	0.754 0.795 0.705
	MI	0.506 0.501 0.514	**0.585** 0.599 0.557	**0.525** 0.509 0.514	**0.585** 0.599 0.557	0.501 0.504 0.498

Table 11: AUROC (error prediction) on AmbigQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best AUROC for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.886 0.894 0.876	**0.887** 0.894 0.876	0.886 0.893 0.876	**0.888** 0.895 0.877	0.886 0.893 0.876
	CoCoA	0.874 0.881 0.862	0.874 0.881 0.864	0.874 0.881 0.862	0.875 0.881 0.863	0.874 0.881 0.862
	Semantic energy	0.865 0.872 0.853	**0.866** 0.873 0.855	0.865 0.872 0.853	**0.867** 0.874 0.856	0.865 0.872 0.853
	Semantic entropy	0.868 0.874 0.858	**0.870** 0.876 0.860	0.868 0.874 0.858	**0.871** 0.876 0.861	0.868 0.874 0.858
	SAR	0.873 0.880 0.861	**0.874** 0.881 0.862	0.873 0.880 0.861	**0.874** 0.881 0.863	0.873 0.880 0.861
	MSP	0.847 0.855 0.835	**0.850** 0.857 0.838	0.848 0.855 0.835	**0.852** 0.860 0.840	0.847 0.855 0.835
	MI	0.839 0.843 0.829	**0.844** 0.848 0.835	0.839 0.842 0.829	**0.848** 0.851 0.840	0.838 0.842 0.829
Qwen3-14B	all UQ	0.861 0.863 0.855	**0.864** 0.866 0.859	0.861 0.863 0.855	**0.872** 0.873 0.867	0.860 0.863 0.854
	CoCoA	0.840 0.842 0.833	**0.848** 0.850 0.841	0.840 0.842 0.833	**0.854** 0.856 0.847	0.840 0.842 0.833
	Semantic energy	0.846 0.849 0.840	**0.853** 0.854 0.847	**0.847** 0.849 0.840	**0.861** 0.861 0.854	0.846 0.849 0.840
	Semantic entropy	0.814 0.815 0.809	**0.847** 0.848 0.841	**0.820** 0.816 0.809	**0.849** 0.849 0.843	0.814 0.816 0.809
	SAR	0.843 0.845 0.838	**0.855** 0.855 0.849	**0.845** 0.845 0.839	**0.862** 0.862 0.856	0.843 0.844 0.838
	MSP	0.828 0.830 0.821	**0.836** 0.838 0.831	**0.828** 0.830 0.821	**0.844** 0.846 0.836	0.828 0.830 0.821
	MI	0.542 0.551 0.534	**0.737** 0.734 0.726	**0.588** 0.551 0.534	**0.737** 0.734 0.726	0.542 0.551 0.534
Spectrum-Llama-3.1-8B	all UQ	0.877 0.886 0.864	**0.878** 0.886 0.865	0.877 0.886 0.864	**0.880** 0.889 0.867	0.877 0.886 0.864
	CoCoA	0.859 0.868 0.846	**0.861** 0.870 0.848	**0.860** 0.868 0.846	**0.863** 0.872 0.849	0.859 0.868 0.846
	Semantic energy	0.857 0.861 0.849	**0.860** 0.864 0.852	0.858 0.861 0.849	**0.863** 0.867 0.854	0.857 0.861 0.849
	Semantic entropy	0.860 0.866 0.851	**0.864** 0.870 0.855	**0.861** 0.866 0.851	**0.867** 0.873 0.857	0.860 0.866 0.851
	SAR	0.864 0.871 0.852	**0.865** 0.873 0.854	0.864 0.872 0.851	**0.867** 0.875 0.855	0.864 0.872 0.851
	MSP	0.820 0.828 0.808	**0.827** 0.836 0.814	**0.821** 0.828 0.807	**0.834** 0.843 0.820	0.820 0.828 0.807
	MI	0.825 0.826 0.818	**0.834** 0.836 0.826	0.823 0.826 0.815	**0.838** 0.841 0.829	0.824 0.826 0.817
Spectrum-Qwen3-14B	all UQ	0.895 0.899 0.888	0.896 0.899 0.889	0.895 0.899 0.888	**0.897** 0.901 0.890	0.895 0.899 0.888
	CoCoA	0.884 0.888 0.876	**0.886** 0.890 0.877	0.884 0.888 0.876	**0.886** 0.892 0.878	0.884 0.888 0.876
	Semantic energy	0.877 0.880 0.870	**0.878** 0.882 0.871	0.877 0.880 0.870	**0.880** 0.884 0.872	0.877 0.880 0.870
	Semantic entropy	0.881 0.884 0.874	**0.884** 0.888 0.876	**0.882** 0.885 0.875	**0.885** 0.889 0.878	**0.882** 0.885 0.875
	SAR	0.886 0.888 0.881	**0.887** 0.889 0.882	0.886 0.888 0.881	**0.888** 0.891 0.882	0.886 0.888 0.881
	MSP	0.854 0.858 0.845	**0.857** 0.861 0.848	0.854 0.858 0.845	**0.860** 0.866 0.850	0.853 0.858 0.845
	MI	0.855 0.852 0.854	**0.861** 0.857 0.859	0.854 0.852 0.854	**0.863** 0.860 0.860	0.855 0.852 0.854
A^2Search-7B	all UQ	0.845 0.855 0.832	**0.848** 0.859 0.834	0.846 0.856 0.832	**0.855** 0.865 0.841	0.845 0.856 0.832
	CoCoA	0.821 0.833 0.806	**0.833** 0.844 0.818	**0.822** 0.833 0.806	**0.835** 0.847 0.818	0.821 0.833 0.806
	Semantic energy	0.838 0.851 0.823	**0.842** 0.856 0.826	**0.840** 0.852 0.823	**0.850** 0.862 0.833	0.838 0.851 0.823
	Semantic entropy	0.815 0.826 0.802	**0.839** 0.850 0.823	**0.820** 0.825 0.802	**0.840** 0.851 0.825	0.815 0.825 0.802
	SAR	0.830 0.840 0.818	**0.840** 0.850 0.827	0.830 0.840 0.818	**0.845** 0.855 0.832	0.830 0.840 0.818
	MSP	0.810 0.824 0.794	**0.819** 0.832 0.802	**0.812** 0.824 0.794	**0.828** 0.840 0.810	0.810 0.824 0.794
	MI	0.553 0.554 0.552	**0.698** 0.699 0.688	**0.588** 0.554 0.552	**0.700** 0.701 0.688	0.553 0.554 0.552

Table 12: AUROC (error prediction) on TriviaQA (validation split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). 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^2Search-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 13: AUROC (error prediction) on NCQA (test split), AlignScore supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best AUROC for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.696 0.731 0.652	0.695 0.728 0.651	0.701 0.734 0.650	0.681 0.709 0.640	0.681 0.727 0.651
	CoCoA	0.644 0.683 0.592	**0.651** 0.688 0.601	0.649 0.684 0.592	0.646 0.680 0.596	0.626 0.683 0.592
	Semantic energy	0.618 0.652 0.571	0.620 0.654 0.573	0.620 0.651 0.571	0.611 0.640 0.567	0.600 0.651 0.571
	Semantic entropy	0.614 0.651 0.557	**0.618** 0.654 0.567	0.616 0.651 0.557	0.604 0.639 0.553	0.592 0.651 0.557
	SAR	0.683 0.722 0.635	0.685 0.721 0.638	**0.690** 0.722 0.635	0.680 0.711 0.634	0.677 0.722 0.635
	MSP	0.584 0.616 0.545	**0.588** 0.617 0.551	**0.588** 0.616 0.545	0.573 0.599 0.536	0.571 0.616 0.545
	MI	0.234 0.179 0.210	**0.241** 0.211 0.249	**0.253** 0.169 0.212	0.230 0.199 0.234	**0.236** 0.173 0.231
Qwen3-14B	all UQ	0.579 0.628 0.525	0.584 0.623 0.545	**0.603** 0.648 0.518	0.564 0.592 0.514	0.579 0.631 0.529
	CoCoA	0.537 0.585 0.476	0.526 0.572 0.452	**0.542** 0.584 0.476	0.483 0.514 0.420	0.520 0.584 0.476
	Semantic energy	0.431 0.483 0.373	**0.478** 0.525 0.399	**0.484** 0.506 0.405	**0.467** 0.503 0.394	0.456 0.483 0.372
	Semantic entropy	0.431 0.488 0.401	**0.472** 0.498 0.421	**0.478** 0.488 0.401	**0.444** 0.463 0.391	0.459 0.488 0.401
	SAR	0.550 0.608 0.481	0.542 0.599 0.473	**0.568** 0.607 0.481	0.525 0.553 0.469	0.563 0.607 0.481
	MSP	0.480 0.526 0.420	**0.482** 0.535 0.410	0.486 0.527 0.417	0.449 0.494 0.371	0.459 0.528 0.417
	MI	0.029 -0.012 0.026	0.030 -0.007 0.008	**0.102** -0.006 0.026	0.030 -0.007 0.008	**0.099** -0.012 -0.004
Spectrum-Llama-3.1-8B	all UQ	0.602 0.621 0.573	0.605 0.621 0.576	0.607 0.623 0.575	**0.607** 0.623 0.582	0.574 0.614 0.570
	CoCoA	0.567 0.579 0.549	**0.572** 0.577 0.559	**0.578** 0.589 0.554	0.572 0.577 0.559	0.550 0.579 0.549
	Semantic energy	0.555 0.571 0.528	**0.559** 0.567 0.542	**0.561** 0.571 0.528	0.537 0.540 0.532	0.519 0.571 0.528
	Semantic entropy	0.557 0.577 0.525	0.559 0.570 0.538	0.559 0.577 0.525	0.531 0.540 0.522	0.510 0.577 0.525
	SAR	0.619 0.632 0.597	**0.621** 0.629 0.601	**0.622** 0.632 0.596	**0.621** 0.629 0.600	0.606 0.632 0.596
	MSP	0.524 0.532 0.509	**0.531** 0.534 0.516	0.526 0.531 0.509	**0.530** 0.532 0.520	0.506 0.531 0.509
	MI	0.194 0.215 0.218	**0.276** 0.246 0.298	**0.243** 0.209 0.216	**0.273** 0.246 0.295	**0.215** 0.207 0.206
Spectrum-Qwen3-14B	all UQ	0.696 0.720 0.660	0.698 0.723 0.659	**0.701** 0.723 0.661	0.688 0.715 0.643	0.675 0.717 0.657
	CoCoA	0.671 0.692 0.641	0.672 0.693 0.640	0.672 0.692 0.641	0.657 0.680 0.622	0.650 0.692 0.640
	Semantic energy	0.645 0.676 0.599	0.647 0.677 0.598	0.649 0.676 0.599	0.636 0.668 0.586	0.624 0.676 0.599
	Semantic entropy	0.655 0.685 0.618	**0.658** 0.685 0.615	**0.659** 0.685 0.618	0.648 0.675 0.607	0.630 0.685 0.618
	SAR	0.684 0.713 0.640	**0.686** 0.715 0.636	**0.689** 0.713 0.640	0.681 0.713 0.629	0.669 0.713 0.640
	MSP	0.628 0.642 0.604	0.626 0.641 0.602	0.627 0.641 0.604	0.611 0.628 0.586	0.605 0.641 0.604
	MI	0.294 0.208 0.260	0.283 0.261 0.293	0.287 0.208 0.255	0.278 0.258 0.284	0.276 0.206 0.240
A^2Search-7B	all UQ	0.593 0.590 0.598	**0.614** 0.617 0.606	**0.599** 0.600 0.596	**0.616** 0.620 0.603	0.568 0.590 0.590
	CoCoA	0.559 0.572 0.548	0.564 0.566 0.554	**0.566** 0.573 0.545	0.551 0.550 0.542	0.529 0.572 0.545
	Semantic energy	0.499 0.533 0.477	**0.556** 0.572 0.528	**0.521** 0.533 0.477	**0.552** 0.570 0.523	0.502 0.532 0.479
	Semantic entropy	0.532 0.524 0.507	**0.553** 0.555 0.543	**0.547** 0.511 0.496	**0.544** 0.550 0.528	0.506 0.511 0.496
	SAR	0.566 0.580 0.555	**0.587** 0.592 0.573	**0.571** 0.579 0.554	**0.581** 0.583 0.571	0.545 0.579 0.554
	MSP	0.511 0.534 0.493	**0.534** 0.549 0.508	**0.522** 0.534 0.491	**0.520** 0.538 0.492	0.485 0.534 0.491
	MI	0.013 -0.015 0.033	**0.161** 0.201 0.067	**0.063** 0.019 -0.019	**0.162** 0.201 0.067	**0.060** 0.019 -0.010

Table 14: PRR on AmbigQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.878	0.878	0.878	0.879	0.869
		0.891	0.892	0.890	0.892	0.891
		0.861	0.860	0.860	0.860	0.860
	CoCoA	0.850	**0.851**	0.850	**0.851**	0.842
		0.860	0.861	0.860	0.862	0.860
		0.834	0.835	0.833	0.834	0.833
	Semantic energy	0.835	**0.838**	0.836	**0.839**	0.825
		0.848	0.852	0.848	0.852	0.848
		0.815	0.818	0.815	0.819	0.815
	Semantic entropy	0.841	**0.846**	0.844	**0.847**	0.835
		0.854	0.861	0.854	0.862	0.854
		0.824	0.825	0.823	0.825	0.823
	SAR	0.861	**0.862**	0.861	0.862	0.852
		0.874	0.874	0.874	0.875	0.874
		0.842	0.843	0.842	0.842	0.842
	MSP	0.808	**0.812**	0.808	**0.816**	0.798
		0.815	0.820	0.815	0.827	0.815
		0.793	0.796	0.793	0.797	0.793
	MI	0.809	**0.826**	0.814	**0.829**	0.802
		0.817	0.839	0.815	0.840	0.815
		0.790	0.807	0.788	0.810	0.790
Qwen3-14B	all UQ	0.793	**0.803**	0.796	**0.821**	0.785
		0.813	0.819	0.813	0.832	0.812
		0.768	0.780	0.768	0.803	0.766
	CoCoA	0.752	**0.770**	0.753	**0.780**	0.744
		0.767	0.778	0.767	0.785	0.767
		0.731	0.754	0.730	0.764	0.731
	Semantic energy	0.755	**0.774**	**0.761**	**0.790**	0.751
		0.782	0.793	0.783	0.803	0.782
		0.722	0.747	0.718	0.767	0.720
	Semantic entropy	0.656	**0.761**	**0.684**	**0.762**	0.678
		0.676	0.765	0.663	0.766	0.678
		0.631	0.746	0.626	0.747	0.626
	SAR	0.744	**0.781**	0.743	**0.786**	0.744
		0.758	0.782	0.758	0.786	0.760
		0.723	0.770	0.723	0.776	0.725
	MSP	0.735	**0.757**	**0.739**	**0.768**	0.729
		0.752	0.765	0.751	0.776	0.751
		0.711	0.741	0.711	0.750	0.711
	MI	0.212	**0.584**	0.193	**0.585**	0.193
		-0.047	0.572	-0.047	0.573	-0.047
		-0.006	0.575	-0.008	0.575	-0.008
Spectrum-Llama-3.1-8B	all UQ	0.851	**0.853**	0.852	**0.856**	0.843
		0.871	0.872	0.871	0.874	0.870
		0.828	0.830	0.827	0.833	0.827
	CoCoA	0.810	**0.814**	0.810	**0.819**	0.798
		0.823	0.828	0.823	0.833	0.823
		0.790	0.794	0.790	0.798	0.790
	Semantic energy	0.812	**0.821**	0.814	**0.823**	0.805
		0.827	0.834	0.827	0.838	0.827
		0.792	0.800	0.792	0.802	0.792
	Semantic entropy	0.819	**0.828**	0.822	**0.830**	0.812
		0.836	0.844	0.835	0.846	0.835
		0.798	0.806	0.798	0.809	0.798
	SAR	0.835	**0.840**	0.835	**0.842**	0.826
		0.849	0.855	0.848	0.857	0.848
		0.815	0.820	0.815	0.821	0.815
	MSP	0.739	**0.750**	0.741	**0.759**	0.731
		0.751	0.760	0.750	0.771	0.750
		0.721	0.731	0.721	0.739	0.721
	MI	0.779	**0.795**	0.768	**0.796**	0.750
		0.788	0.805	0.787	0.807	0.787
		0.761	0.778	0.734	0.778	0.761
Spectrum-Qwen3-14B	all UQ	0.871	**0.872**	0.871	**0.876**	0.864
		0.883	0.885	0.883	0.890	0.883
		0.852	0.854	0.852	0.856	0.852
	CoCoA	0.846	**0.849**	0.847	**0.849**	0.838
		0.855	0.859	0.855	0.860	0.855
		0.829	0.831	0.829	0.832	0.829
	Semantic energy	0.835	**0.839**	0.836	**0.841**	0.829
		0.850	0.853	0.850	0.856	0.850
		0.815	0.818	0.815	0.819	0.815
	Semantic entropy	0.845	**0.848**	**0.846**	**0.850**	0.839
		0.856	0.859	0.856	0.862	0.856
		0.822	0.829	0.821	0.830	0.821
	SAR	0.862	**0.865**	0.863	**0.865**	0.854
		0.871	0.874	0.871	0.875	0.871
		0.846	0.848	0.846	0.848	0.846
	MSP	0.793	**0.804**	0.796	**0.811**	0.785
		0.801	0.814	0.800	0.824	0.800
		0.778	0.786	0.778	0.790	0.778
	MI	0.816	**0.830**	0.801	**0.829**	0.802
		0.831	0.837	0.818	0.837	0.831
		0.800	0.815	0.781	0.813	0.797
A^2Search-7B	all UQ	0.736	**0.748**	0.742	**0.769**	0.734
		0.755	0.769	0.759	0.783	0.755
		0.711	0.721	0.707	0.746	0.703
	CoCoA	0.672	**0.725**	**0.680**	**0.731**	0.670
		0.694	0.741	0.694	0.749	0.694
		0.643	0.699	0.643	0.700	0.643
	Semantic energy	0.710	0.734	**0.720**	**0.749**	0.713
		0.733	0.756	0.736	0.772	0.741
		0.681	0.701	0.681	0.715	0.676
	Semantic entropy	0.655	**0.738**	**0.662**	**0.740**	0.658
		0.675	0.751	0.675	0.753	0.675
		0.630	0.715	0.630	0.716	0.630
	SAR	0.696	**0.740**	**0.698**	**0.750**	0.692
		0.712	0.755	0.712	0.764	0.712
		0.673	0.717	0.673	0.725	0.673
	MSP	0.652	**0.703**	**0.663**	**0.718**	0.654
		0.676	0.720	0.676	0.735	0.676
		0.621	0.676	0.620	0.688	0.620
	MI	0.222	**0.524**	0.179	**0.531**	0.186
		-0.022	0.528	-0.009	0.536	-0.009
		0.031	0.499	0.026	0.503	0.026

Table 15: PRR on TriviaQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). 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^2Search-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 16: PRR on NCQA (test split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best PRR for model.

Model	Features	baseline	latent gated	oracle gated	latent selective	oracle selective
Llama-3.1-8B	all UQ	0.799 0.812 0.784	0.798 0.810 0.783	0.799 0.812 0.783	0.797 0.808 0.784	0.797 0.810 0.783
	CoCoA	0.768 0.783 0.751	**0.771** 0.786 0.754	0.769 0.783 0.751	0.767 0.783 0.750	0.768 0.783 0.751
	Semantic energy	0.752 0.765 0.736	0.753 0.766 0.737	0.753 0.764 0.737	0.750 0.764 0.736	0.752 0.764 0.737
	Semantic entropy	0.754 0.769 0.736	**0.756** 0.771 0.738	0.754 0.768 0.736	0.754 0.771 0.736	0.754 0.768 0.736
	SAR	0.789 0.804 0.773	0.789 0.804 0.773	**0.790** 0.803 0.773	0.788 0.804 0.770	0.789 0.803 0.773
	MSP	0.735 0.748 0.719	0.735 0.747 0.721	0.735 0.748 0.719	0.735 0.749 0.720	0.735 0.748 0.719
	MI	0.581 0.576 0.585	**0.592** 0.578 0.594	**0.598** 0.582 0.584	0.569 0.573 0.576	0.582 0.579 0.584
Qwen3-14B	all UQ	0.752 0.768 0.739	0.756 0.773 0.740	**0.758** 0.773 0.738	0.753 0.768 0.740	0.754 0.769 0.741
	CoCoA	0.719 0.730 0.706	0.718 0.730 0.704	**0.721** 0.729 0.706	0.712 0.727 0.701	0.719 0.729 0.707
	Semantic energy	0.699 0.713 0.685	**0.706** 0.721 0.689	**0.707** 0.716 0.687	0.701 0.720 0.681	0.699 0.713 0.685
	Semantic entropy	0.699 0.710 0.687	**0.705** 0.715 0.692	0.703 0.710 0.686	0.692 0.711 0.681	0.699 0.710 0.687
	SAR	0.743 0.762 0.725	0.736 0.752 0.719	**0.748** 0.762 0.726	0.741 0.764 0.722	0.743 0.762 0.726
	MSP	0.687 0.700 0.672	**0.691** 0.707 0.674	0.688 0.700 0.670	0.675 0.698 0.661	0.686 0.700 0.670
	MI	0.505 0.501 0.509	0.506 0.474 0.517	**0.537** 0.502 0.509	0.506 0.525 0.494	0.505 0.501 0.509
Spectrum-Llama-3.1-8B	all UQ	0.748 0.748 0.749	0.749 0.748 0.750	0.750 0.750 0.749	0.748 0.747 0.749	0.746 0.745 0.747
	CoCoA	0.727 0.724 0.731	**0.730** 0.725 0.734	**0.735** 0.734 0.735	0.726 0.723 0.729	0.727 0.724 0.731
	Semantic energy	0.721 0.719 0.724	**0.723** 0.719 0.726	**0.724** 0.719 0.724	0.722 0.721 0.724	0.721 0.719 0.724
	Semantic entropy	0.724 0.726 0.722	0.725 0.726 0.723	0.726 0.726 0.722	0.724 0.725 0.722	0.725 0.726 0.722
	SAR	0.755 0.753 0.758	**0.757** 0.755 0.758	**0.756** 0.753 0.758	0.752 0.751 0.756	0.755 0.753 0.758
	MSP	0.702 0.697 0.706	**0.704** 0.698 0.709	0.703 0.696 0.706	**0.703** 0.698 0.708	0.701 0.696 0.706
	MI	0.578 0.569 0.588	**0.605** 0.588 0.619	**0.589** 0.569 0.589	0.554 0.542 0.572	0.577 0.565 0.590
Spectrum-Qwen3-14B	all UQ	0.803 0.810 0.795	0.804 0.811 0.796	**0.804** 0.811 0.796	0.804 0.812 0.794	0.802 0.809 0.793
	CoCoA	0.789 0.795 0.781	0.789 0.795 0.780	0.789 0.795 0.781	0.788 0.795 0.780	0.789 0.795 0.781
	Semantic energy	0.774 0.783 0.763	0.775 0.783 0.763	0.774 0.783 0.762	0.775 0.784 0.764	0.774 0.783 0.762
	Semantic entropy	0.778 0.787 0.766	**0.779** 0.788 0.769	0.778 0.787 0.766	0.775 0.783 0.768	0.778 0.787 0.766
	SAR	0.795 0.802 0.785	**0.796** 0.803 0.786	**0.796** 0.801 0.785	0.793 0.800 0.784	0.795 0.801 0.785
	MSP	0.763 0.768 0.757	0.763 0.767 0.756	0.763 0.767 0.757	0.760 0.763 0.756	0.763 0.767 0.757
	MI	0.581 0.561 0.605	**0.600** 0.590 0.610	**0.591** 0.565 0.604	**0.595** 0.589 0.602	0.579 0.565 0.605
A^2Search-7B	all UQ	0.757 0.748 0.772	**0.767** 0.760 0.776	**0.764** 0.754 0.773	**0.768** 0.763 0.776	0.756 0.748 0.771
	CoCoA	0.736 0.729 0.746	0.739 0.731 0.747	**0.740** 0.730 0.746	0.734 0.727 0.744	0.736 0.729 0.746
	Semantic energy	0.724 0.724 0.728	**0.732** 0.731 0.733	0.729 0.724 0.728	0.725 0.723 0.730	0.724 0.723 0.728
	Semantic entropy	0.727 0.724 0.731	**0.731** 0.726 0.735	**0.731** 0.724 0.730	**0.731** 0.726 0.735	0.727 0.724 0.730
	SAR	0.744 0.739 0.756	**0.756** 0.752 0.761	**0.749** 0.738 0.756	**0.756** 0.750 0.761	0.744 0.739 0.756
	MSP	0.708 0.708 0.711	0.713 0.712 0.712	0.712 0.708 0.711	**0.712** 0.711 0.713	0.708 0.708 0.711
	MI	0.502 0.495 0.509	**0.555** 0.577 0.524	**0.526** 0.506 0.503	**0.555** 0.577 0.524	0.502 0.506 0.497

Table 17: AUROC (error prediction) on AmbigQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). 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	0.903	0.904	0.903
		0.914	0.914	0.913	0.914	0.913
		0.891	0.891	0.890	0.891	0.890
	CoCoA	0.884	**0.885**	0.884	**0.885**	0.884
		0.893	0.893	0.892	0.894	0.892
		0.872	0.872	0.871	0.872	0.871
	Semantic energy	0.874	**0.875**	0.874	**0.876**	0.874
		0.885	0.887	0.885	0.887	0.885
		0.858	0.860	0.858	0.861	0.858
	Semantic entropy	0.880	**0.881**	0.880	**0.882**	0.880
		0.891	0.893	0.891	0.893	0.891
		0.865	0.866	0.865	0.867	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	**0.856**	0.853	**0.858**	0.853
		0.858	0.862	0.858	0.865	0.858
		0.843	0.845	0.843	0.846	0.843
	MI	0.863	**0.869**	0.864	**0.871**	0.863
		0.872	0.878	0.872	0.879	0.872
		0.850	0.855	0.850	0.857	0.850
Qwen3-14B	all UQ	0.860	**0.863**	0.860	**0.872**	0.859
		0.869	0.872	0.869	0.878	0.869
		0.848	0.852	0.848	0.861	0.847
	CoCoA	0.835	**0.841**	0.835	**0.847**	0.834
		0.842	0.847	0.841	0.851	0.841
		0.824	0.831	0.823	0.836	0.824
	Semantic energy	0.840	**0.846**	**0.842**	**0.853**	0.840
		0.852	0.856	0.853	0.861	0.852
		0.826	0.832	0.826	0.839	0.826
	Semantic entropy	0.811	**0.841**	**0.818**	**0.842**	0.810
		0.818	0.846	0.818	0.847	0.819
		0.801	0.829	0.799	0.831	0.799
	SAR	0.842	**0.855**	0.842	**0.858**	0.843
		0.850	0.859	0.850	0.862	0.851
		0.832	0.845	0.832	0.849	0.833
	MSP	0.820	**0.829**	**0.822**	**0.835**	0.820
		0.828	0.835	0.828	0.841	0.828
		0.809	0.818	0.809	0.823	0.809
	MI	0.538	**0.735**	**0.585**	**0.735**	0.538
		0.545	0.731	0.545	0.731	0.545
		0.533	0.726	0.532	0.726	0.532
Spectrum-Llama-3.1-8B	all UQ	0.884	**0.885**	0.885	**0.887**	0.884
		0.896	0.897	0.896	0.898	0.895
		0.870	0.871	0.870	0.872	0.870
	CoCoA	0.857	**0.860**	0.857	**0.862**	0.857
		0.866	0.869	0.866	0.872	0.866
		0.845	0.847	0.845	0.849	0.845
	Semantic energy	0.858	**0.862**	0.859	**0.863**	0.858
		0.867	0.871	0.867	0.873	0.867
		0.845	0.850	0.845	0.850	0.845
	Semantic entropy	0.864	**0.867**	0.864	**0.869**	0.864
		0.874	0.877	0.874	0.879	0.874
		0.851	0.854	0.850	0.856	0.850
	SAR	0.875	0.876	0.875	**0.877**	0.875
		0.884	0.885	0.884	0.886	0.884
		0.863	0.864	0.862	0.865	0.862
	MSP	0.811	**0.818**	0.812	**0.823**	0.811
		0.817	0.823	0.816	0.829	0.816
		0.802	0.808	0.802	0.812	0.802
	MI	0.843	**0.849**	0.840	**0.850**	0.842
		0.849	0.855	0.849	0.857	0.849
		0.832	0.838	0.829	0.839	0.832
Spectrum-Qwen3-14B	all UQ	0.902	**0.903**	0.902	**0.904**	0.902
		0.912	0.913	0.912	0.915	0.912
		0.889	0.890	0.888	0.891	0.889
	CoCoA	0.887	**0.888**	0.887	**0.888**	0.887
		0.895	0.896	0.895	0.896	0.895
		0.875	0.876	0.875	0.876	0.875
	Semantic energy	0.877	**0.879**	0.877	**0.880**	0.877
		0.887	0.889	0.887	0.890	0.887
		0.864	0.866	0.864	0.867	0.864
	Semantic entropy	0.884	**0.886**	**0.884**	**0.887**	**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	**0.897**	0.895	**0.897**	0.895
		0.904	0.905	0.904	0.906	0.904
		0.883	0.884	0.883	0.884	0.883
	MSP	0.852	**0.857**	0.853	**0.861**	0.852
		0.857	0.863	0.857	0.868	0.857
		0.843	0.847	0.843	0.849	0.843
	MI	0.871	**0.876**	0.867	**0.877**	0.871
		0.879	0.883	0.878	0.884	0.879
		0.860	0.865	0.856	0.865	0.860
A^2Search-7B	all UQ	0.838	**0.842**	0.839	**0.849**	0.836
		0.845	0.849	0.845	0.855	0.845
		0.829	0.832	0.827	0.840	0.826
	CoCoA	0.808	**0.822**	**0.810**	**0.825**	0.808
		0.816	0.829	0.816	0.832	0.816
		0.797	0.811	0.797	0.812	0.797
	Semantic energy	0.824	**0.830**	**0.827**	**0.837**	0.825
		0.834	0.839	0.834	0.846	0.834
		0.814	0.818	0.814	0.825	0.814
	Semantic entropy	0.804	**0.828**	**0.809**	**0.830**	0.804
		0.810	0.834	0.810	0.835	0.810
		0.796	0.818	0.796	0.819	0.796
	SAR	0.822	**0.833**	**0.823**	**0.839**	0.822
		0.829	0.839	0.829	0.844	0.829
		0.813	0.824	0.813	0.828	0.813
	MSP	0.794	**0.805**	**0.797**	**0.814**	0.794
		0.803	0.813	0.803	0.822	0.803
		0.783	0.793	0.783	0.800	0.783
	MI	0.545	**0.709**	**0.585**	**0.715**	0.545
		0.545	0.708	0.547	0.714	0.547
		0.546	0.700	0.546	0.703	0.546

Table 18: AUROC (error prediction) on TriviaQA (validation split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). 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^2Search-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 19: AUROC (error prediction) on NCQA (test split), LLM-judge supervision. **Bold** = significant improvement over baseline ($p < 0.05$, test set). Highlight = best AUROC for model.