Evidential Semantic Uncertainty Decomposition for Large Language Models

Kangshuo Li[1]*, Tianhao Wang[1]*, Yifan Zhang[1], Lance M. Kaplan[2]
Audun Jøsang[3], Dong Hyun Jeong[4], Jin-Hee Cho[5], Feng Chen[1]
[1]Department of Computer Science, The University of Texas at Dallas
[2]US Army Research Laboratory [3]University of Oslo
[4]University of the District of Columbia [5]Virginia Tech
{kangshuo.li,tianhao.wang,yifan.zhang3,feng.chen}@utdallas.edu
lkaplan@ieee.org, audun.josang@mn.uio.no, djeong@udc.edu, jicho@vt.edu

Abstract

Free-form language generation poses a challenge for uncertainty quantification because responses do not belong to a fixed global label space. Semantic uncertainty methods address this by measuring the semantic diversity of sampled responses. Yet semantic diversity alone cannot provide a principled decomposition into aleatoric uncertainty associated with conflicting plausible meanings and epistemic uncertainty associated with unsupported guessing. Moreover, raw cluster counts depend on the sampling budget and should not be mistaken for evidential strength. We propose an evidential probing framework (EPN) for semantic uncertainty decomposition. For each prompt, sampled responses define a semantic frame of distinct output semantics and provide initial support for each semantic class. A lightweight evidence-retention probe learns how much of this support should be retained as reliable semantic evidence. The retained evidence induces a subjective opinion in which dissonance captures conflict among supported semantic alternatives and vacuity captures insufficient retained evidence. To train EPN, we introduce Augmented Semantic Uncertainty Cross-Entropy (AS-UCE), an LLM-tailored evidential objective that augments the prompt-specific semantic frame with an unsupported class, allowing unreliable generation support to be routed away from observed semantic clusters. Extensive empirical results suggest that EPN with AS-UCE achieves average absolute AUROC gains of 4.3 and 1.9 percentage points for out-of-distribution (OOD) and ambiguity detection, respectively, over the strongest task-relevant baseline, with better task specialization, while requiring fewer sampled responses than perturbation-based decomposition methods. Code is available at: https://github.com/kl3259/llm-evidential-uncertainty-decomposition.

1 Introduction

Large language models (LLMs) have become increasingly capable of free-form generation across question answering, reasoning, and agentic usage. Despite their capability, they can still produce responses that are factually incorrect, semantically inconsistent, or unsupported by available evidence. This gap between linguistic plausibility and reliability makes uncertainty quantification indispensable to trustworthy LLM deployment. Existing methods estimate uncertainty from token probabilities, self-reported confidence, response consistency, hidden representations, or semantic similarity among sampled generations [1, 3, 5, 11, 13, 18, 20, 23, 27, 31, 36, 40]. Among them, semantic uncertainty methods are particularly relevant for free-form generation, where they sample multiple responses,

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

cluster them according to semantic equivalence, and estimate uncertainty over prompt-specific semantic classes [6, 20, 35]. This representation is important because LLM outputs do not belong to a fixed global label space. The same answer may appear in many semantically equivalent surface forms, and different prompts may induce different sets of plausible meanings.

However, measuring total uncertainty is only the first step toward understanding unreliable generation. Following a widely used conceptual decomposition [7, 8, 15, 22], predictive uncertainty can be split into aleatoric (data) uncertainty and epistemic (model) uncertainty. Epistemic uncertainty arises when the correct output is predictable in principle, but the model has limited training evidence, representation, or capacity for it. In contrast, aleatoric uncertainty arises from ambiguity, under-specification, or inherent variability in the input or data-generating process. For example, a question such as "*Who is the governor of this state?*" is ambiguous without specifying the state and time of reference. This distinction guides interventions. High epistemic uncertainty points to retrieval, tool use, further supervision, or deferring to a human, whereas high aleatoric uncertainty points to asking for the missing context or keeping several valid answers.

Semantic dispersion can reflect genuine ambiguity or insufficient support that leads to unstable or hallucinated guesses. A scalar semantic uncertainty score cannot tell these two sources apart, and therefore cannot indicate which of these actions is appropriate. Moreover, raw semantic cluster counts depend on the chosen sampling budget and should not be treated as intrinsic evidential strength. Existing decomposition methods often rely on input clarification [12] or prompt perturbations [28]. Their effectiveness depends on the quality and coverage of these perturbations, while additional generations increase inference cost. We therefore seek a decomposition that works directly with the semantic frame of a prompt, the set of distinct output semantics among its sampled responses. Unlike a fixed label space, both its members and its size vary with the prompt. Such a representation should distinguish supported ambiguity from insufficient support and account for unsupported responses during evidence retention.

We propose a prompt-level evidential probing framework (EPN) for fine-grained semantic uncertainty decomposition in LLM generation. Our design separates semantic allocation under a user-specified sampling budget from reliable evidential support. For each prompt, sampled responses define a prompt-specific semantic frame, and length-normalized backbone likelihoods provide initial response support, reflecting the model's confidence in generating each response. A lightweight probe learns how much of this support to retain as semantic evidence. *Semantic allocation* is the distribution of sampled responses over the semantic frame, and *evidence retention* stands for the fraction of their initial support the probe keeps as reliable semantic evidence. Two prompts with the same allocation can therefore differ in uncertainty when their retained evidence differs. The retained evidence forms a subjective opinion in which vacuity captures insufficient retained evidence and dissonance captures conflict among supported semantic alternatives. To train the probe, we introduce Augmented Semantic Uncertainty Cross-Entropy (AS-UCE), an evidential objective designed for free-form LLM generation. Standard uncertainty cross-entropy assumes a fixed label space and assigns all target mass to known classes, which is unsuitable when semantic classes are prompt-specific, and some generated responses are unsupported. Our theoretical analysis shows that retaining unsupported evidence can distort the decomposition, while filtering it out leaves the amount of rejected evidence unrepresented in the semantic opinion. AS-UCE augments the prompt-specific semantic frame with an unsupported class, explicitly representing and supervising evidence rejection during training. Our work makes the following contributions:

- We formulate EPN, a framework that separates semantic allocation from evidence retention to construct subjective opinions over prompt-specific semantic classes.
- We introduce AS-UCE to supervise both retained and unsupported evidence, with theoretical analysis explaining its role in semantic uncertainty decomposition. We study the optional pseudo-OOD extension, Reg-AS-UCE, separately as an auxiliary-supervision ablation.
- We empirically show that EPN trained with AS-UCE achieves average absolute AUROC gains of 4.3 percentage points for OOD detection and 1.9 percentage points for ambiguity detection over the strongest task-relevant baseline. Additional analyses show better overall task specialization and weaker average association between the two components than uncertainty decomposition baselines.
- EPN requires fewer sampled responses than perturbation-based decomposition methods, while EPN-IT further reduces the response sampling budget.

2 Related Work

(1) Uncertainty quantification for LLMs. LLM uncertainty has been estimated from likelihood, consistency, self-evaluation, and representations. Likelihood-based methods use token-level entropy [14] or length-normalized entropy [30]. Semantic methods compare sampled meanings, including semantic entropy over equivalence classes [6, 20] and Kernel Language Entropy with soft semantic relationships [35]. Other methods use self-evaluation, such as P(True) [18] and confidence verbalization [18, 24, 43], or response consistency, such as SelfCheckGPT [31]. Representation-based methods use geometric response-embedding scores [3, 25] or lightweight probes for truthfulness, semantic uncertainty, and ambiguity [1, 10, 19, 36, 47]. These approaches provide useful reliability scores, but typically estimate total uncertainty rather than separating semantic ambiguity from insufficient support. **(2) Uncertainty decomposition for LLMs.** Recent decomposition methods often rely on auxiliary variables that clarify, perturb, or augment the context. Input clarification ensembling estimates ambiguity-driven uncertainty from valid clarifications and treats residual uncertainty as model-side uncertainty [12]; ensemble-of-ensembles uses semantically perturbed prompts as proxy posterior samples and decomposes within- and across-prompt variation in embedding space [28]. In in-context learning, uncertainty has been decomposed by varying demonstrations and model configurations [26], or by using auxiliary queries to derive variational bounds on aleatoric and epistemic uncertainty [16]. Another line attributes uncertainty from multiple responses to ambiguity, knowledge gaps, or both [34]. These methods are valuable, but depend on the quality and coverage of external interventions; our method instead induces the decomposition from a subjective opinion over prompt-specific semantic classes. **(3) Evidential and Bayesian semantic uncertainty.** Bayesian and evidential variants improve semantic entropy estimation under limited sampling. Bayesian semantic entropy places a Dirichlet prior over latent semantic distributions and conditions on samples and sequence probabilities [4], while Evidential Semantic Entropy assigns Dempster–Shafer mass to observed clusters and residual mass to the full frame for unobserved or underspecified answers [21]. These methods estimate total semantic uncertainty. In contrast, we learn how much generated support should be retained as semantic evidence, yielding vacuity for insufficient support and dissonance for conflict among retained semantic alternatives.

3 Preliminaries

Evidential neural networks. Evidential neural networks [38] (ENNs) quantify predictive uncertainty by outputting a distribution over categorical probabilities rather than a single probability vector. For a K-class classification problem, an ENN predicts nonnegative evidence $\mathbf{e} = [e_1, \ldots, e_K]$, where $e_k \geq 0$, and constructs Dirichlet parameters $\alpha_k = e_k + 1$, with total strength $\alpha_0 = \sum_{k=1}^{K} \alpha_k = K + \sum_{k=1}^{K} e_k$. The Dirichlet distribution $\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ represents a second-order distribution over class probability vectors, whose predictive mean is $\mathbb{E}[\mathbf{p}] = \boldsymbol{\alpha}/\alpha_0$. From the subjective-logic perspective, e_k is the evidence supporting class k. The corresponding belief mass, vacuity, and dissonance are

$$b_k = \frac{e_k}{K + \sum_{j=1}^{K} e_j}, \qquad u = \frac{K}{K + \sum_{j=1}^{K} e_j}, \qquad D(\mathbf{b}) = \sum_{i=1}^{K} b_i \frac{\sum_{j \neq i} b_j \mathrm{Bal}(b_j, b_i)}{\sum_{j \neq i} b_j}. \quad (1)$$

Here, the summand in $D(\mathbf{b})$ is defined as zero when $\sum_{j \neq i} b_j = 0$, and $\mathrm{Bal}(b_j, b_i) = 1 - \frac{|b_j - b_i|}{b_j + b_i}$ when $b_i + b_j > 0$, and 0 otherwise. Vacuity is high when the total evidence is small, while dissonance is high when multiple classes receive substantial and balanced belief.

Uncertainty cross-entropy (UCE). Given a target distribution $\mathbf{y} = [y_1, \ldots, y_K]$, UCE trains the Dirichlet opinion by minimizing the expected cross-entropy under the predicted Dirichlet: $\mathcal{L}_{\mathrm{UCE}} = \mathbb{E}_{\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\alpha})} \big[- \sum_{k=1}^{K} y_k \log p_k \big]$, where $\mathbf{p}$ is a random probability vector on the $K-1$-dimensional simplex. Using $\mathbb{E}_{\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\alpha})}[\log p_k] = \psi(\alpha_k) - \psi(\alpha_0)$, where $\psi(\cdot)$ is the digamma function, this can be rewritten as

$$\mathcal{L}_{\mathrm{UCE}} = \sum_{k=1}^{K} y_k \left[\psi(\alpha_0) - \psi(\alpha_k) \right]. \quad (2)$$

Standard UCE assumes a fixed label space and assigns all target mass to known classes. This assumption is natural for closed-set classification, but restrictive for free-form LLM generation, where the semantic answer space is prompt-specific, and some generated responses may be unsupported or

implausible rather than valid low-probability members of the observed semantic frame. Our method adapts evidential learning to this setting by constructing prompt-specific Dirichlet opinions and augmenting UCE with an unsupported class.

4 Methodology

We propose a prompt-level evidential probing framework (EPN) for decomposing semantic uncertainty in free-form LLM generation. The method addresses two challenges. First, unlike closed-set classification, the semantic answer space is prompt-specific and must be induced from sampled responses. Second, raw cluster counts should not be interpreted as intrinsic evidential strength, since they depend on the sampling budget. Inspired by [46], our framework separates *semantic allocation* from *evidence retention*: sampled responses define the semantic frame, response likelihoods provide initial support for semantic classes, and a lightweight prompt-level probe learns how much support should be retained as reliable semantic evidence.

4.1 Prompt-Specific Semantic Frame

Let x be an input prompt and p_θ be a frozen backbone LLM. We sample M responses $\mathcal{R}(x) = \{r^{(1)}, \ldots, r^{(M)}\}$ and group them into $K = K(x)$ prompt-specific semantic frame $\mathcal{C}(x) = \{c_1, \ldots, c_K\}$. Let $a(r^{(m)}) \in \{1, \ldots, K\}$ denote the semantic assignment of response $r^{(m)}$. We define $\mathcal{R}_k(x) = \{r^{(m)} : a(r^{(m)}) = k\}$ as the responses assigned to c_k. For each response $r^{(m)} = (y_1^{(m)}, \ldots, y_{T_m}^{(m)})$, we compute its length-normalized backbone likelihood

$$s_{\text{norm}}^{(m)}(x) = \exp\left(\frac{1}{T_m}\sum\nolimits_{t=1}^{T_m} \log p_\theta(y_t^{(m)} \mid x, y_{<t}^{(m)})\right), \tag{3}$$

which serves as initial response support and reflects how strongly the backbone favors each response among its own samples, without being a calibrated probability of correctness. The total initial support is $S_{\text{init}}(x) = \sum_{m=1}^{M} s_{\text{norm}}^{(m)}(x)$. Since $s_{\text{norm}}^{(m)}(x) \le 1$, the sampling budget M bounds $S_{\text{init}}(x)$, and the probe introduced next decides how much of it to retain.

4.2 Prompt-Level Evidence Retention

Let $\mathbf{h}(x) \in \mathbb{R}^d$ denote a prompt representation extracted from the frozen backbone. The evidential probe $g_\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}$ predicts a discount score $\rho(x) = \sigma(g_\phi(\mathbf{h}(x))) \in (0, 1)$ using the sigmoid function $\sigma(\cdot)$. We convert this score into the retention factor

$$q(x) = \exp(-\zeta\rho(x)), \;\; q(x) \in (e^{-\zeta}, 1), \tag{4}$$

where $\zeta > 0$ controls the discount strength. The factor $q(x)$ is shared by all sampled responses for the same prompt. In other words, the probe does not classify individual responses; instead, it estimates how much prompt-level generated support should be retained as semantic evidence. The retained evidence for semantic class c_k is defined as

$$e_k(x) = \sum\nolimits_{r^{(m)} \in \mathcal{R}_k(x)} q(x) s_{\text{norm}}^{(m)}(x), \;\; \forall k \in [K]. \tag{5}$$

The total retained evidence is $S_{\text{ret}}(x) = \sum_{k=1}^{K} e_k(x) = q(x) S_{\text{init}}(x)$. This design reflects the intended role of vacuity. When the backbone has weak support for answering the prompt, all sampled answers should be discounted, even if they form coherent semantic clusters.

4.3 Augmented Semantic Uncertainty Cross-Entropy (AS-UCE)

Standard UCE assumes that all target mass belongs to observed classes. This assumption is unsuitable for LLM generation because some sampled responses provide unreliable support for the prompt-specific semantic answer classes. We therefore introduce AS-UCE, which augments the prompt-specific semantic frame with an unsupported class $c_\perp$. This class is not an output semantic at inference time but an evidential rejection component used during training. The evidence routed to $c_\perp$ can come from responses of different kinds, such as hallucinated answers that contradict the reference and

guesses the backbone cannot support. AS-UCE groups them by role. Both withhold support from the observed semantic classes, so the loss sends their evidence to $c_\perp$ instead of letting it inflate any class. In practice, we estimate this rejected mass from answer correctness judgments, so hallucinated and otherwise incorrect responses serve as a proxy for unsupported evidence.

During training, initial support is split into retained semantic evidence and unsupported evidence. Specifically, $e_\perp(x) = (1 - q(x))S_{\text{init}}(x)$, and the augmented Dirichlet parameters are given by

$$\boldsymbol{\alpha}^+(x) = \mathbf{e}^+(x) + \mathbf{1}_{K+1}, \text{where } \mathbf{e}^+(x) = [e_1(x), \ldots, e_K(x), e_\perp(x)]. \tag{6}$$

Let $\mathbf{y}^+(x) = [y_1^+(x), \ldots, y_K^+(x), y_\perp^+(x)]$ be a soft semantic target distribution over the augmented semantic frame $\Omega^+(x) = \{c_1, \ldots, c_K, c_\perp\}$. An auxiliary probe trained on answer correctness judgments estimates a shared reliability score for each prompt. We weight each semantic class's initial support by this score, assign the rejected support to $c_\perp$, and normalize these masses to obtain $\mathbf{y}^+(x)$ (see Appendix D.3). AS-UCE is the expected cross-entropy between this augmented target and a categorical distribution whose probability vector $\mathbf{p}^+ = [p_1^+, \ldots, p_{K(x)}^+, p_\perp^+]$ is sampled from the augmented Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha}^+(x))$:

$$\mathcal{L}_{\text{AS-UCE}}(x) = \mathbb{E}_{\mathbf{p}^+ \sim \text{Dir}(\boldsymbol{\alpha}^+(x))} \left[-\sum\nolimits_{k=1}^{K} y_k^+(x) \log p_k^+ - y_\perp^+(x) \log p_\perp^+ \right]. \tag{7}$$

Using the closed-form Dirichlet expectation of log probabilities, this becomes

$$\mathcal{L}_{\text{AS-UCE}}(x) = \sum\nolimits_{k=1}^{K} y_k^+(x) \big[\psi(\alpha_0^+(x)) - \psi(\alpha_k^+(x)) \big] + y_\perp^+(x) \big[\psi(\alpha_0^+(x)) - \psi(\alpha_\perp^+(x)) \big], \tag{8}$$

where $\alpha_0^+(x) = \sum_{k=1}^{K} \alpha_k^+(x) + \alpha_\perp^+(x)$. If target mass is assigned to semantic classes, AS-UCE encourages evidence retention, if target mass is assigned to $c_\perp$, it encourages routing initial support away from the semantic frame. The loss trains the shared retention factor without changing the relative allocation of initial support across semantic classes. Optionally, AS-UCE can be combined with a pseudo-OOD regularizer used in ENNs [29, 41], which pulls the retained opinion of pseudo-OOD prompts toward a flat Dirichlet prior. We refer to this variant as Reg-AS-UCE (see Appendix C).

4.4 Inference and Uncertainty Decomposition

At inference time, the unsupported class is not treated as a semantic answer. Given an input x, we sample responses, construct $\mathcal{C}(x)$, compute $s_{\text{norm}}^{(m)}(x)$, evaluate $q(x)$, and form retained semantic evidence $\mathbf{e}(x) = [e_1(x), \ldots, e_K(x)]$. The inference-time Dirichlet total strength is given by $\alpha_0(x) = K + S_{\text{ret}}(x)$. The corresponding belief masses and vacuity are

$$b_k(x) = \frac{e_k(x)}{K + S_{\text{ret}}(x)}, \qquad u(x) = \frac{K}{K + S_{\text{ret}}(x)}. \tag{9}$$

We use vacuity to estimate epistemic uncertainty, $U_{\text{epi}}(x) = u(x)$, which is high when retained semantic evidence is small, and subjective-logic dissonance to estimate aleatoric uncertainty, $U_{\text{ale}}(x) = D(\mathbf{b}(x))$, which is high when multiple semantic classes receive substantial and balanced belief mass. Following prior analyses of loss-trained second-order models [2, 39], we take this usage of vacuity as an operational hypothesis. Since aleatoric and epistemic estimates often remain correlated in practice [33], Section 6 and Appendix G.4 test the hypothesis directly through cross-evaluation of both components on both tasks, a controlled intervention that varies ambiguity and supplied knowledge separately, and the measured dependence between the two components.

In low-budget settings, we optionally use an iterative prompting variant, denoted as EPN-IT. Instead of drawing a fixed set of independent responses, EPN-IT sequentially prompts the backbone to produce semantically distinct responses, while allowing it to output `NONE` when no additional plausible answer is found. The accepted non-`NONE` responses are treated as proxy semantic classes, avoiding an additional clustering step and aiming to limit redundant or implausible additions to the semantic frame. Implementation details of EPN-IT are provided in Appendix E.

5 Theoretical Analysis

We analyze how evidence retention shapes semantic uncertainty decomposition and how AS-UCE learns this retention from reliability supervision. A sampled semantic frame describes the meanings

expressed by generated responses, while the retained evidence controls the evidential strength assigned to them. Standard semantic UCE supervises the observed semantic classes, so its treatment of unsupported generations depends on how evidence and targets are constructed. We first show how counting rejected evidence in the semantic subjective opinion affects vacuity and dissonance, and then establish how introducing an unsupported class into the AS-UCE target supplies separate retention and rejection gradient terms, connecting reliability supervision to the amount of evidence entering that opinion.

Theorem 1 (Unfiltered semantic UCE can distort the uncertainty decomposition)**.** *Fix a prompt x and a semantic frame with $K \geq 2$ classes. Let $A_k(x) = \sum_{r^{(m)} \in \mathcal{R}_k(x)} s^{(m)}_{\mathrm{norm}}(x)$ denote the unfiltered initial support for semantic class k, with total support $S_{\mathrm{init}}(x) = \sum_{k=1}^{K} A_k(x) > 0$. For $0 < q(x) < 1$, the retained evidence vector $\mathbf{e}(x)$ and rejected evidence vector $\boldsymbol{\delta}(x)$ have components $e_k(x) = q(x)A_k(x)$ and $\delta_k(x) = (1 - q(x))A_k(x)$ for $k \in [K]$. Counting the rejected evidence in its original semantic classes yields*

$$u(\mathbf{e}(x) + \boldsymbol{\delta}(x)) < u(\mathbf{e}(x)), \;\; D(\mathbf{b}(\mathbf{e}(x) + \boldsymbol{\delta}(x))) \geq D(\mathbf{b}(\mathbf{e}(x))). \tag{10}$$

The dissonance inequality is strict whenever at least two semantic classes have positive unfiltered initial support.

Proof sketch. Since $e_k(x) + \delta_k(x) = A_k(x)$ and $S_{\mathrm{ret}}(x) = q(x)S_{\mathrm{init}}(x) < S_{\mathrm{init}}(x)$, the vacuity inequality follows directly. For the fixed prompt, define the belief scaling factor $c = \frac{K+S_{\mathrm{ret}}(x)}{q(x)(K+S_{\mathrm{init}}(x))} > 1$. The induced belief vectors satisfy $\mathbf{b}(\mathbf{e}(x) + \boldsymbol{\delta}(x)) = c\,\mathbf{b}(\mathbf{e}(x))$. Common scaling leaves each balance term and the normalized weights of competing classes unchanged, so $D(c\mathbf{b}(\mathbf{e}(x))) = cD(\mathbf{b}(\mathbf{e}(x)))$. If at least two classes have positive initial support, the retained opinion has positive dissonance, giving a strict increase. A full proof is provided in Appendix B.2. □

When the rejected evidence is indeed unsupported, counting it as semantic evidence lowers vacuity and amplifies existing dissonance, potentially making unsupported diversity appear as supported semantic conflict. The relative allocation across semantic classes stays unchanged, but the additional evidence increases their total belief mass and magnifies existing conflict. Filtering out this evidence removes its contribution, yet the retained subjective opinion alone does not record how much was rejected, as shown by Theorem 3 in Appendix B.4. If filtering is paired with fixed zero semantic target weights for a fully unsupported example, the UCE expression also yields no learning signal. These observations motivate explicitly representing rejected evidence during training and assigning it target mass, so that reliability supervision can guide how much initial support to retain.

Theorem 2 (AS-UCE provides a gradient signal for evidence rejection)**.** *For a prompt x with $S_{\mathrm{init}}(x) > 0$, hold the semantic frame, initial support, and augmented target $\mathbf{y}^{+}(x)$ fixed. The AS-UCE loss satisfies*

$$\frac{\partial \mathcal{L}_{\text{AS-UCE}}(x)}{\partial q(x)} = -\sum_{k=1}^{K} y^{+}_{k}(x)A_k(x)\psi'\big(\alpha^{+}_{k}(x)\big) + y^{+}_{\perp}(x)S_{\mathrm{init}}(x)\psi'\big(\alpha^{+}_{\perp}(x)\big), \tag{11}$$

where ψ' is the trigamma function. The second term provides an evidence-rejection signal that is strictly positive whenever $y^{+}_{\perp}(x) > 0$. When it exceeds the magnitude of the semantic-retention term, decreasing $q(x)$ is a descent direction, reducing retained semantic evidence and increasing vacuity. In particular, $y^{+}_{\perp}(x) = 1$ guarantees $\frac{\partial \mathcal{L}_{\text{AS-UCE}}(x)}{\partial q(x)} > 0$.

Proof sketch. The augmented total strength $\alpha^{+}_{0}(x) = K + 1 + S_{\mathrm{init}}(x)$ is independent of $q(x)$. Differentiating AS-UCE with the initial support and $\mathbf{y}^{+}(x)$ held fixed gives Eq. 11. The sign conclusions follow from $\psi'(z) > 0$ for $z > 0$. The first term encourages retaining evidence for semantic classes with positive target mass, while the second term is positive whenever unsupported target mass exists. Thus, when the rejection term dominates, a sufficiently small feasible decrease in $q(x)$ lowers the loss. Since $u(x) = K/(K+q(x)S_{\mathrm{init}}(x))$ is strictly decreasing in $q(x)$, reducing $q(x)$ increases vacuity. Moreover, the semantic belief masses $b_k(x) = q(x)A_k(x)/(K + q(x)S_{\mathrm{init}}(x))$ shrink as unsupported evidence is rejected, reducing retained belief conflict that could otherwise cause inflated dissonance. A full proof is provided in Appendix B.3. □

Theorem 2 explains how AS-UCE turns the augmented soft reliability target into a direct learning signal for the evidence-retaining factor $q(x)$. Its explicit rejection coordinate contributes a distinct gradient term, allowing the probe to learn how much generated support to retain as semantic evidence. For fixed initial support, changing $q(x)$ leaves the augmented total strength $\alpha_0^+(x) = K+1+S_{\text{init}}(x)$ unchanged while adjusting the retained semantic strength $K + S_{\text{ret}}(x)$. This redistributes the same initial support between the semantic frame and the rejection component. The relative sizes of the retention and rejection terms determine the local adjustment favored by the loss. When rejection dominates, a feasible decrease in retention raises vacuity and attenuates existing dissonance. AS-UCE thus connects reliability supervision to the distinction between insufficient retained evidence and conflict among supported semantic alternatives.

6 Experiments

6.1 Empirical Setup

Our experiments target the following research questions: (1) whether our proposed evidential probing with AS-UCE is effective at estimating decomposed multidimensional uncertainties; (2) whether EPN achieves better decomposition of aleatoric and epistemic uncertainty, with lower component dependence and better task specialization; (3) whether efficiency-driven EPN-IT achieves comparable performance to EPN with a smaller inference-time sampling budget. We evaluate the proposed EPN on OOD and ambiguity detection tasks for factual question answering with free-form generation.

OOD detection. We adopt the setting that uses a test-time held-out set of unknown questions as OOD. For in-distribution samples, we use `TriviaQA` [17]. For OOD samples, we use `Alcuna` [45] containing attributive questions about synthetic novel animals, and a constructed `TemporalOOD` dataset of factual questions of news and events after November 2025 to ensure it is beyond all backbones' knowledge cutoff. We combine `TriviaQA` with each OOD held-out set to create two OOD detection benchmarks: `Alcuna-TriviaQA` and `TemporalOOD-TriviaQA`. Pseudo-OODs can be added to the training and validation splits as a regularization signal, provided that they come from a different domain than the genuine test OOD source and are also unknown. Since `Alcuna` and `TemporalOOD` cover different topic scopes, they can serve as pseudo-OOD sources for each other. Details of the dataset construction are shown in Appendix D.1. We use vacuity $U_{\text{epi}}(x)$ as the predicted uncertainty score for the OOD detection task.

Ambiguity detection. Here, ambiguity refers to underspecified questions that admit multiple valid interpretations and can therefore lead to semantically distinct but plausible answers. We use `AmbigQA` [32], a benchmark where each question is annotated as either `multipleQAs` for ambiguous questions or `singleAnswer` for clear questions. We apply the ambiguity label as $y_i^{\text{amb}} = \mathbb{I}(y_i = \texttt{multipleQAs})$, and use predicted dissonance $U_{\text{ale}}(x)$ as the uncertainty score for this task.

Baselines. We primarily consider methods that explicitly attempt to decompose uncertainty in free-form generation, including **Input Clarification Ensembling (ICE)** [12] and **Ensemble-of-Ensembles (EoE)** [28], which use clarified or perturbed inputs to expose variability for the decomposition. For self-consistency methods, we compare against **Semantic Entropy (SE)** [20], **SelfCheckGPT** [31], **Kernel Language Entropy (KLE)** [35], **EigenValue (EigV)** [25], and **Semantic Density (SD)** [37]. We also consider the hybrid uncertainty quantification approach **CoCoA** [42], which utilizes confidence and self-consistency. Details of baselines can be found in Appendix D.2.

Evaluation settings. We evaluate the proposed method on a diverse set of instruction-tuned white-box LLMs with different scales from different model families, including Llama-3.2-3B-Instruct, Meta-Llama-3-8B-Instruct, Mistral-Small-24B-Instruct-2501, and Qwen3-30B-A3B-Instruct-2507 [9, 44]. For both OOD detection and ambiguity detection, we report threshold-free discrimination metrics, including the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPR). We evaluate uncertainty decomposition validity by visualizing the Euclidean distance from each method's mean $(\text{AUROC}_{U_{\text{ale}}}, \text{AUROC}_{U_{\text{epi}}})$ pair to the ideal point $(0.5, 1.0)$ for OOD detection and $(1.0, 0.5)$ for ambiguity detection. Implementation details are provided in Appendix D.3.

Table 1: OOD and ambiguity detection results from a single run. EPN is trained with AS-UCE and uses vacuity for OOD detection and dissonance for ambiguity detection. Blue and light blue cells mark the best and runner-up values in each row.

Dataset	Backbone	SE		SelfCheckGPT		ICE		EoE		EPN w/ AS-UCE	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Alcuna-TriviaQA	Llama-3.2-3B	0.800	0.703	0.770	0.653	0.763	**0.753**	0.649	0.580	**0.817**	0.694
	Llama-3-8B	0.875	0.854	0.851	0.834	0.856	**0.860**	0.709	0.617	**0.897**	0.855
	Mistral-Small-24B	0.919	0.914	0.899	0.866	0.866	0.870	0.768	0.650	**0.944**	**0.930**
	Qwen3-30B-A3B	0.751	0.704	0.721	0.691	0.755	0.698	0.607	0.526	**0.879**	**0.806**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.697	0.587	0.656	0.575	0.695	0.669	0.735	0.634	**0.876**	**0.776**
	Llama-3-8B	0.789	0.722	0.782	0.729	0.770	**0.748**	0.782	0.693	**0.829**	**0.748**
	Mistral-Small-24B	0.852	0.832	0.849	0.813	0.835	0.808	0.788	0.679	**0.895**	**0.872**
	Qwen3-30B-A3B	0.754	0.721	0.751	0.727	0.770	0.763	0.662	0.564	**0.854**	**0.806**
AmbigQA	Llama-3.2-3B	0.459	0.524	0.461	0.530	0.490	0.502	0.426	0.488	**0.528**	**0.565**
	Llama-3-8B	0.476	0.521	0.480	0.515	0.502	0.527	0.441	0.479	**0.533**	**0.570**
	Mistral-Small-24B	0.475	0.516	0.495	0.518	0.494	0.511	0.473	0.497	**0.533**	**0.609**
	Qwen3-30B-A3B	0.499	0.525	0.482	0.507	0.502	0.508	0.465	0.494	**0.553**	**0.606**

Table 2: Uncertainty decomposition validity results. Each cell reports the mean ± sample standard deviation of AUROC over five data-split seeds. EPN is trained with AS-UCE, and its dissonance (D) and vacuity (V) are the aleatoric (AU) and epistemic (EU) uncertainty components. Blue and light blue cells mark the best and runner-up means in each row, where ICE, EoE, and EPN enter with EU for OOD detection and with AU for ambiguity detection.

Dataset	Backbone	SE	SelfCheck GPT	KLE	EigV	CoCoA		SD	ICE		EoE		EPN w/ AS-UCE	
						MaxProb	PPL		AU	EU	AU	EU	D (AU)	V (EU)
Alcuna-TriviaQA	Llama-3.2-3B	0.796 ±0.010	0.767 ±0.008	0.697 ±0.020	0.683 ±0.018	0.646 ±0.013	0.595 ±0.015	0.609 ±0.016	0.637 ±0.016	0.802 ±0.016	0.652 ±0.006	0.643 ±0.023	0.502 ±0.015	**0.813 ±0.007**
	Llama-3-8B	0.864 ±0.008	0.841 ±0.006	0.863 ±0.007	0.853 ±0.008	0.845 ±0.010	0.803 ±0.009	0.827 ±0.009	0.535 ±0.020	0.832 ±0.014	0.710 ±0.004	0.676 ±0.017	0.654 ±0.010	**0.880 ±0.016**
	Mistral-Small-24B	0.916 ±0.003	0.901 ±0.004	0.912 ±0.004	0.916 ±0.003	0.925 ±0.002	0.919 ±0.002	0.909 ±0.002	0.720 ±0.122	0.902 ±0.020	0.851 ±0.005	0.771 ±0.021	0.765 ±0.035	**0.936 ±0.016**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.680 ±0.014	0.624 ±0.019	0.580 ±0.012	0.558 ±0.013	0.587 ±0.016	0.539 ±0.017	0.546 ±0.016	0.311 ±0.011	0.693 ±0.010	0.794 ±0.005	0.728 ±0.009	0.457 ±0.011	**0.858 ±0.006**
	Llama-3-8B	0.790 ±0.004	0.780 ±0.004	0.797 ±0.003	0.800 ±0.004	0.793 ±0.007	0.755 ±0.013	0.796 ±0.008	0.682 ±0.013	0.757 ±0.012	0.823 ±0.002	0.768 ±0.020	0.658 ±0.017	**0.855 ±0.007**
	Mistral-Small-24B	0.852 ±0.002	0.852 ±0.004	0.851 ±0.003	0.857 ±0.004	0.860 ±0.002	0.847 ±0.001	0.851 ±0.003	0.626 ±0.034	0.847 ±0.007	0.888 ±0.003	0.792 ±0.014	0.731 ±0.025	**0.897 ±0.015**
AmbigQA	Llama-3.2-3B	0.474 ±0.009	0.478 ±0.011	0.459 ±0.007	0.459 ±0.008	0.464 ±0.011	0.469 ±0.014	0.457 ±0.012	0.509 ±0.025	0.516 ±0.033	0.459 ±0.024	0.494 ±0.020	**0.535 ±0.025**	0.469 ±0.010
	Llama-3-8B	0.510 ±0.013	0.500 ±0.013	0.503 ±0.015	0.512 ±0.015	0.506 ±0.011	0.506 ±0.014	0.506 ±0.010	0.503 ±0.006	0.498 ±0.020	0.484 ±0.027	0.515 ±0.030	**0.530 ±0.009**	0.513 ±0.013
	Mistral-Small-24B	0.514 ±0.025	0.518 ±0.016	0.509 ±0.021	0.515 ±0.021	0.527 ±0.021	0.521 ±0.020	0.521 ±0.022	0.500 ±0.017	0.471 ±0.002	0.516 ±0.016	0.527 ±0.007	**0.540 ±0.016**	0.512 ±0.018

6.2 Main Results

OOD and ambiguity detection. Tables 1 and 2 show consistent AUROC gains for EPN using vacuity for OOD detection and dissonance for ambiguity detection. We measure every gain against the strongest baseline in the same setting, where decomposition baselines contribute their epistemic score for OOD detection and their aleatoric score for ambiguity detection. Across the four backbones in Table 1, EPN trained with AS-UCE improves AUROC by 6.2 and 4.0 percentage points on average for OOD and ambiguity detection. Table 2 extends the comparison to modern UQ baselines and reports results over five data-split seeds for three backbones. EPN again achieves the highest mean AUROC in all nine settings, with average gains of 4.3 and 1.9 percentage points. However, the absolute AUROC values on AmbigQA remain close to 0.5, especially for smaller backbones, indicating that ambiguity detection is substantially harder than OOD detection. This difficulty is not unique to our method, since ICE, which explicitly targets input ambiguity through clarification, also degrades in this setting. One explanation is that some models respond to annotated ambiguous questions as if they lacked knowledge; in such cases, epistemic uncertainty can dominate the response behavior,

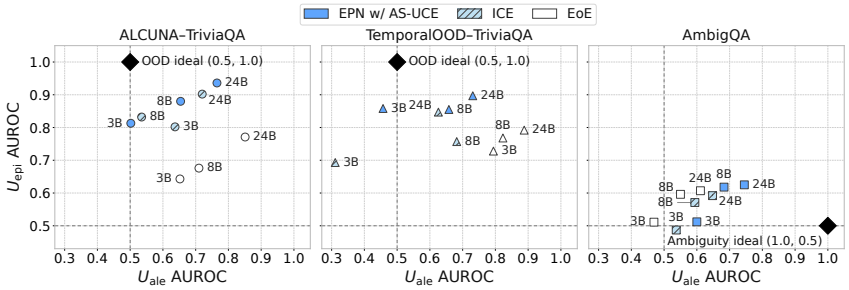


Figure 1: Cross-evaluation of aleatoric and epistemic uncertainty measures. Points show mean AUROC over five data-split seeds. Black diamonds mark the task-specific ideal, and points closer to it indicate better task specialization. The AmbigQA panels use only the retained questions.

Table 3: Inference sampling efficiency results from a single run. Each cell reports AUROC, and positive $\Delta\text{AUROC}_{\text{IT}}$ favors EPN. EPN and EPN-IT are both trained with Reg-AS-UCE for OOD detection and with AS-UCE for ambiguity detection.

Method	Alcuna-TriviaQA				TemporalOOD-TriviaQA				AmbigQA			
	3B	8B	24B	30B	3B	8B	24B	30B	3B	8B	24B	30B
$\text{AUROC}_{\text{EPN}}$	0.835	0.905	0.963	0.902	0.889	0.831	0.899	0.862	0.528	0.533	0.533	0.553
$\text{AUROC}_{\text{EPN-IT}}$	0.829	0.918	0.971	0.916	0.902	0.810	0.901	0.864	0.497	0.499	0.523	0.535
$\Delta\text{AUROC}_{\text{IT}}$	0.006	-0.013	-0.008	-0.014	-0.013	0.020	-0.002	-0.002	0.031	0.033	0.011	0.017

weakening the semantic conflict signal captured by dissonance. Appendix F provides diagnostic evidence and case studies illustrating this uncertainty overriding behavior.

Uncertainty decomposition validity. We now examine task specialization and the dependence between U_{ale} and U_{epi}. In Table 2, EPN's vacuity is the stronger OOD score and its dissonance is the stronger ambiguity score, while the other component stays closer to chance in all nine settings. The two measures therefore respond to different uncertainty patterns. Given the mismatch between human-annotated ambiguity and the backbone's perceived ambiguity, we additionally evaluate all uncertainty decomposition methods on a shared AmbigQA subset, excluding annotated-ambiguous questions with only one output semantic cluster. Across the OOD test sets and retained AmbigQA questions, Figure 1 shows that EPN is closest to the task-specific ideal in six of nine settings and achieves the lowest mean distance, which is 16.8% below ICE and 40.1% below EoE. Additional component dependence analyses and controlled experiments in Appendix G.4 further support EPN's uncertainty decomposition. Across the nine settings, EPN reduces the mean absolute Spearman correlation by 45.7% and 54.2% relative to ICE and EoE, respectively.

Inference sampling efficiency. A subjective opinion enables uncertainty decomposition without Bayesian decomposition over clarified or perturbed inputs [12, 28]. ICE and EoE require $O(N_{\text{pert}}M_{\text{per}})$ generations for N_{pert} clarified or perturbed inputs with M_{per} samples each, whereas EPN requires only $O(M)$ generations from the original prompt. EPN uses $M = 10$, reducing the response-sampling budget while preserving a direct decomposition signal. To compare EPN to the sampling-efficient variant EPN-IT, we measure $\Delta\text{AUROC}_{\text{IT}} = \text{AUROC}_{\text{EPN}} - \text{AUROC}_{\text{EPN-IT}}$. Table 3 shows that EPN-IT preserves most of the task-relevant AUROC while using up to $M = 5$ responses. Across the four backbones, $|\Delta\text{AUROC}_{\text{IT}}|$ stays within 0.014 on ALCUNA, 0.020 on TemporalOOD, and 0.033 on AmbigQA, which makes both EPN and EPN-IT sampling efficient options for uncertainty decomposition.

6.3 Ablation Study

Pseudo-OOD regularization. Optional pseudo-OOD regularization can further improve OOD detection when its trade-off coefficient λ is properly chosen. With $\lambda = 0.25$, Reg-AS-UCE raises vacuity AUROC over AS-UCE on both OOD datasets for all four backbones, by 1.2 percentage points on average. Appendix G.5 reports the full comparison and a sweep over λ.

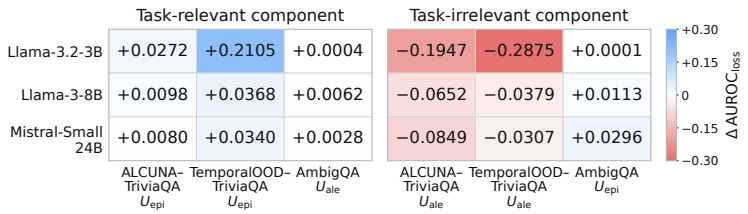


Figure 2: Mean paired $\Delta\text{AUROC}_{\text{loss}} = \text{AUROC}_{\text{AS-UCE}} - \text{AUROC}_{\text{UCE}}$ across five random seeds. Panels separate the task-relevant and task-irrelevant uncertainty components, identified below each dataset. Blue and red indicate increases and decreases, respectively.

Comparison between AS-UCE and UCE. We compare UCE and AS-UCE across five paired seeds with matched data splits and training configurations. Figure 2 shows that AS-UCE improves mean task-relevant AUROC in all nine backbone–dataset settings, with an average gain of 5.44 percentage points in OOD detection. Notably, across all six OOD settings, AS-UCE improves vacuity AUROC and brings dissonance AUROC closer to chance level (0.5), reducing the latter from 0.751 to 0.634 on average. Thus, AS-UCE strengthens vacuity's OOD utility while simultaneously reducing off-target AUROC in dissonance. This corroborates AS-UCE as a tailored objective for open-domain generation. Its additional unsupported class $c_\perp$ enables direct supervision of evidence rejection beyond the observed semantic classes, which benefits distinguishing insufficient support from conflict among supported semantic alternatives. Appendix G.6 provides additional comparisons with pseudo-OOD regularization.

Additional ablations. In addition, we compare common design choices of aleatoric uncertainty measures in Appendix G.7. Appendix G.8 examines EPN robustness to correctness judges of different capacities and noisy supervision, where EPN remains effective with smaller and same-backbone judges and under moderate label noise. Appendix G.9 asks whether response-level retention is worth the added complexity. We found that prompt-level EPN performs better in most OOD settings and comparably on ambiguity detection, while requiring fewer probe evaluations. Finally, Appendix G.10 turns to correctness prediction within each OOD source. Vacuity remains informative on TemporalOOD but not ALCUNA, revealing that identifying unfamiliar questions and assessing the correctness of their answers are distinct challenges.

7 Limitations

Our study has several limitations that motivate future work. First, EPN has not yet been extended to longer responses. One route is the factoid-level procedure used for long-form semantic entropy [6], which splits a long answer into short factual claims and asks a question about each one. EPN would then run on each claim, taking the prompt, the answer so far, and the claim-specific question as its input. Richer prompt representations are natural next steps. Second, our implementation requires access to both backbone likelihoods and hidden states, and a fully black-box extension remains open. Third, vacuity is a learned second-order score, and empirical results show that the two uncertainty components still share some variation across backbones. These limitations do not change the main contribution that AS-UCE provides an explicit evidential mechanism for separating retained semantic support from unsupported generated evidence in prompt-specific semantic frames.

8 Conclusion

This work advances semantic uncertainty decomposition for free-form LLM generation. We propose EPN to learn how much generated support to retain, forming prompt-specific subjective opinions in which vacuity captures insufficient support and dissonance captures conflict among supported semantic alternatives. The proposed AS-UCE loss explicitly represents and supervises unsupported evidence during training, preserving information about how much generated support is rejected. Extensive experiments demonstrate that EPN is effective in practice, with improved OOD and ambiguity detection, better overall task specialization, and robustness to weak judges and moderate label noise. The EPN-IT variant further reduces the response sampling budget.

Acknowledgments and Disclosure of Funding

This work is supported by the National Science Foundation (NSF) under Grant Nos. 2417747, 2220574, 2107449, and 2107450.

References

[1] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023.

[2] Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35:29205–29216, 2022.

[3] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Zj12nzlQbz`.

[4] Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy. *arXiv preprint arXiv:2504.03579*, 2025.

[5] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, 2024.

[6] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[8] Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. In *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pages 45–87. Springer, 2020.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Jiatong Han, Jannik Kossen, Muhammed Razzak, and Yarin Gal. Semantic entropy neurons: Encoding semantic uncertainty in the latent space of LLMs. In *MINT: Foundation Model Interventions*, 2024. URL `https://openreview.net/forum?id=BbZKxrZCNn`.

[11] Jianfeng He, Linlin Yu, Changbin Li, Runing Yang, Fanglan Chen, Kangshuo Li, Min Zhang, Shuo Lei, Xuchao Zhang, Mohammad Beigi, et al. Survey of uncertainty estimation in llms-sources, methods, applications, and challenges. *Information Fusion*, page 104057, 2025.

[12] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*, 2023.

[13] Xu Hu, Yifan Zhang, Songtao Wei, Chen Zhao, Qiannan Li, Bingzhe Li, and Feng Chen. Small updates, big doubts: Does parameter-efficient fine-tuning enhance hallucination detection? *arXiv preprint arXiv:2602.11166*, 2026.

[14] Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023.

[15] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[16] I. Shavindra Jayasekera, Jacob Si, Filippo Valdettaro, Wenlong Chen, Aldo A. Faisal, and Yingzhen Li. Variational uncertainty decomposition for in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=MqGZIJxZ1z.

[17] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017.

[18] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[19] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

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

[21] Lucie Kunitomo-Jacquin, Edison Marrese-Taylor, Ken Fukuda, and Masahiro Hamasaki. Evidential semantic entropy for LLM uncertainty quantification. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7107–7122, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.334. URL https://aclanthology.org/2026.eacl-long.334/.

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[23] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, 2021.

[24] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

[25] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

[26] Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, et al. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, 2024.

[27] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117, 2025.

[28] Wang Ma, Debarun Bhattacharjya, Junkyu Lee, Nhan H Pham, Harsha Kokel, and Qiang Ji. Black-box uncertainty quantification for large language models via ensemble-of-ensembles. In *AAAI 2026 Workshop on Assessing and Improving Reliability of Foundation Models in the Real World*, 2026. URL https://openreview.net/forum?id=jfq9DUrW90.

[29] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[30] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

[31] Potsawee Manakul, Adian Liusie, and Mark Gales. 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, 2023.

[32] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 5783–5797, 2020.

[33] Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems*, 37:50972–51038, 2024.

[34] Yang Nan, Pengfei He, Ravi Tandon, and Han Xu. Can multiple responses from an LLM reveal the sources of its uncertainty? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15551–15569, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.841. URL https://aclanthology.org/2025.findings-emnlp.841/.

[35] Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=j2wCrWmgMX.

[36] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*, 2024.

[37] Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533, 2024.

[38] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[39] Maohao Shen, Jongha Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory W Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? *Advances in Neural Information Processing Systems*, 37:107830–107864, 2024.

[40] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38, 2025.

[41] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021.

[42] Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Cocoa: A minimum bayes risk framework bridging confidence and consistency for uncertainty quantification in llms. *Advances in Neural Information Processing Systems*, 38:106236–106281, 2026.

[43] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

[44] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[45] Xunjian Yin, Baizhou Huang, and Xiaojun Wan. Alcuna: Large language models meet new knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1397–1414, 2023.

[46] Linlin Yu, Kangshuo Li, Pritom Kumar Saha, Yifei Lou, and Feng Chen. Evidential uncertainty probes for graph neural networks. *arXiv preprint arXiv:2503.08097*, 2025.

[47] Zhuoxuan Zhang, Jinhao Duan, Edward Kim, and Kaidi Xu. Sparse neurons carry strong signals of question ambiguity in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16081–16099, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.813. URL https://aclanthology.org/2025.emnlp-main.813/.

A Notation

Table 4: Summary of notation.

Notation	Description
$\mathrm{Dir}(\cdot)$	Dirichlet distribution
$\sigma(\cdot)$	Sigmoid function
$\psi(\cdot)$	Digamma function
$g_\phi(\cdot)$	Evidential probe
$\mathbf{h}(x)$	Hidden states from a backbone LLM for input x
x	Input prompt
p_θ	Backbone LLM
M	Number of sampled responses per question
$r^{(m)}$	m-th sampled response
$\mathcal{R}(x)$	Indexed collection of sampled responses for input x
$\mathcal{R}_k(x)$	Collection of sampled responses belonging to the k-th semantic cluster
$y_t^{(m)}$	t-th generated token for response m
T_m	Length of the m-th response
$s_{\text{norm}}^{(m)}(x)$	Length-normalized generated sequence likelihood
$\mathcal{C}(x)$	Semantic clusters for input x
K	Number of semantic clusters
c_k	k-th semantic cluster
$A_k(x)$	Unfiltered initial support for semantic class k
$S_{\text{init}}(x)$	Total initial support
ζ	Discount strength
$\rho(x)$	Probe-predicted discount score
$q(x)$	Evidence-retention factor
$e_k(x)$	Retained evidence for the k-th semantic class
$\delta_k(x)$	Rejected evidence $(1-q(x))A_k(x)$ for semantic class k
$\alpha_k(x)$	Retained strength for the k-th semantic class
$y_k^+(x)$	Semantic reliability target for the k-th semantic class
$S_{\text{ret}}(x)$	Total retained evidence
$c_\perp$	Unsupported semantic class
$e_\perp(x)$	Unsupported evidence
$\alpha_\perp^+(x)$	Strength for the unsupported semantic class
$y_\perp^+(x)$	Semantic target for the unsupported semantic class
$\mathbf{e}(x)$	Retained Dirichlet evidence vector for K observed semantic clusters
$\boldsymbol{\alpha}(x)$	Retained Dirichlet strength vector for K observed semantic clusters
α_0	Total strength
$\mathbf{e}^+(x)$	Augmented evidence vector with an additional unsupported semantic class
$\boldsymbol{\alpha}^+(x)$	Augmented strength vector with an additional unsupported semantic class
$\mathbf{y}^+(x)$	Soft semantic target distribution over the augmented semantic frame
α_0^+	Total augmented strength
$\mathbf{b}(x)$	Belief vector
$u(x)$	Vacuity
$D(\mathbf{b}(x))$	Dissonance
$\mathrm{Bal}(\cdot)$	Balance term used in dissonance
$\Omega^+(x)$	Augmented semantic frame
$\mathcal{L}_{\text{UCE}}$	Uncertainty Cross-Entropy loss
$\mathcal{L}_{\text{AS-UCE}}$	Augmented Semantic Uncertainty Cross-Entropy loss
$\mathcal{L}_{\text{POOD}}$	Pseudo-OOD KL divergence regularizer
$\mathrm{KL}(\cdot\|\cdot)$	KL divergence

B Proofs and Additional Theoretical Analysis

This appendix provides full proofs of the two theoretical results stated in Section 5, followed by additional analysis that complements the main results. Our goal is to clarify how the proposed evidential construction separates retained semantic support from unsupported generated evidence. The proof of Theorem 1 characterizes the distortion that arises when unsupported evidence is retained in the semantic opinion under the stated assumptions, lowering vacuity and amplifying existing

dissonance. Standard UCE can also suppress evidence inconsistent with its target through an implicit gradient signal. The proof of Theorem 2 shows that AS-UCE introduces a direct gradient signal for routing unsupported evidence away from semantic classes. The subsequent dissonance analysis further shows that conflict among retained beliefs diminishes as evidence is rejected, reducing the risk of mistaking unsupported diversity for supported ambiguity. Finally, we describe how KL divergence-based pseudo-OOD regularization encourages high vacuity on auxiliary unsupported prompts by pushing their retained semantic opinions toward a flat Dirichlet prior.

B.1 Setup and Notation

Consider a prompt x with K prompt-specific semantic classes $\mathcal{C}(x) = \{c_1, \ldots, c_K\}$. Let $\mathcal{R}(x) = \{r^{(1)}, \ldots, r^{(M)}\}$ denote the sampled responses. We treat responses as indexed observations, so identical text sampled at different indices is counted separately. Let $a(r^{(m)}) \in \{1, \ldots, K\}$ be the semantic class assignment of response $r^{(m)}$. For each class c_k, define

$$\mathcal{R}_k(x) = \{r^{(m)} \in \mathcal{R}(x) : a(r^{(m)}) = k\}. \tag{12}$$

Each response is assigned a length-normalized backbone likelihood $s^{(m)}_{\text{norm}}(x) \in [0, 1]$, which provides initial response support. The unfiltered initial support $A_k(x)$ for semantic class k and total initial support $S_{\text{init}}(x)$ are defined as

$$A_k(x) = \sum_{r^{(m)} \in \mathcal{R}_k(x)} s^{(m)}_{\text{norm}}(x), \qquad S_{\text{init}}(x) = \sum_{k=1}^{K} A_k(x) = \sum_{m=1}^{M} s^{(m)}_{\text{norm}}(x). \tag{13}$$

The prompt-level evidence-retention probe outputs $q(x) \in (e^{-\zeta}, 1)$. At inference time, the retained semantic evidence is

$$e_k(x) = q(x)A_k(x), \qquad S_{\text{ret}}(x) = \sum_{k=1}^{K} e_k(x) = q(x)S_{\text{init}}(x). \tag{14}$$

The resulting subjective opinion over the retained semantic frame has belief masses and vacuity

$$b_k(x) = \frac{e_k(x)}{K + S_{\text{ret}}(x)}, \qquad u(x) = \frac{K}{K + S_{\text{ret}}(x)}. \tag{15}$$

We use $u(x)$ as epistemic uncertainty and $D(\mathbf{b}(x))$ as aleatoric uncertainty, where $D(\cdot)$ is the multinomial dissonance functional defined in Eq. 1.

During training with AS-UCE, we augment the semantic frame with an unsupported class $c_\perp$. The unsupported evidence is

$$e_\perp(x) = (1 - q(x))S_{\text{init}}(x), \tag{16}$$

and the augmented Dirichlet parameters are

$$\alpha^+_k(x) = 1 + q(x)A_k(x), \qquad k = 1, \ldots, K, \tag{17}$$

and

$$\alpha^+_\perp(x) = 1 + (1 - q(x))S_{\text{init}}(x). \tag{18}$$

The augmented total strength is

$$\alpha^+_0(x) = \sum_{k=1}^{K} \alpha^+_k(x) + \alpha^+_\perp(x) = K + 1 + S_{\text{init}}(x), \tag{19}$$

which is independent of $q(x)$. Thus, the retention probe does not create or destroy initial support; it redistributes initial support between the semantic frame and the unsupported component.

B.2 Full Proof of Theorem 1

Theorem 1 (Unfiltered semantic UCE can distort the uncertainty decomposition)**.** *Fix a prompt x and a semantic frame with $K \geq 2$ classes. Let $A_k(x) = \sum_{r^{(m)} \in \mathcal{R}_k(x)} s^{(m)}_{\text{norm}}(x)$ denote the*

unfiltered initial support for semantic class k, with total support $S_{\text{init}}(x) = \sum_{k=1}^{K} A_k(x) > 0$. For $0 < q(x) < 1$, the retained evidence vector $\mathbf{e}(x)$ and rejected evidence vector $\boldsymbol{\delta}(x)$ have components $e_k(x) = q(x)A_k(x)$ and $\delta_k(x) = (1 - q(x))A_k(x)$ for $k \in [K]$. Counting the rejected evidence in its original semantic classes yields

$$u(\mathbf{e}(x) + \boldsymbol{\delta}(x)) < u(\mathbf{e}(x)), \;\; D(\mathbf{b}(\mathbf{e}(x) + \boldsymbol{\delta}(x))) \geq D(\mathbf{b}(\mathbf{e}(x))). \tag{10}$$

The dissonance inequality is strict whenever at least two semantic classes have positive unfiltered initial support.

Proof. Fix the prompt x and its K-class semantic frame. The unfiltered initial support $A_k(x)$ for each semantic class k splits into retained evidence $e_k(x) = q(x)A_k(x)$ and rejected evidence $\delta_k(x) = (1 - q(x))A_k(x)$. Hence,

$$e_k(x) + \delta_k(x) = A_k(x), \qquad \sum_{k=1}^{K} \delta_k(x) = S_{\text{init}}(x) - S_{\text{ret}}(x) = e_{\perp}(x). \tag{20}$$

The vector $\boldsymbol{\delta}(x)$ therefore records the rejected evidence by its original semantic class. Both opinions use the same K classes. For any nonnegative evidence vector $\mathbf{v} = [v_1, \ldots, v_K]$, its vacuity and belief vector $\mathbf{b}(\mathbf{v}) = [b_1(\mathbf{v}), \ldots, b_K(\mathbf{v})]$ are

$$u(\mathbf{v}) = \frac{K}{K + \sum_{j=1}^{K} v_j}, \qquad b_k(\mathbf{v}) = \frac{v_k}{K + \sum_{j=1}^{K} v_j}, \qquad k \in [K]. \tag{21}$$

We first compare vacuity. Since $0 < q(x) < 1$ and $S_{\text{init}}(x) > 0$, we have $S_{\text{ret}}(x) = q(x)S_{\text{init}}(x) < S_{\text{init}}(x)$. Thus,

$$u(\mathbf{e}(x) + \boldsymbol{\delta}(x)) = \frac{K}{K + S_{\text{init}}(x)} < \frac{K}{K + S_{\text{ret}}(x)} = u(\mathbf{e}(x)). \tag{22}$$

For dissonance, the belief masses induced by the two evidence vectors satisfy

$$b_k(\mathbf{e}(x)) = \frac{q(x)A_k(x)}{K + S_{\text{ret}}(x)}, \qquad b_k(\mathbf{e}(x) + \boldsymbol{\delta}(x)) = \frac{A_k(x)}{K + S_{\text{init}}(x)}. \tag{23}$$

Restoring the rejected evidence changes both the evidence numerator and the total strength in the denominator. To express the resulting belief mass as a multiple of its retained counterpart, factor the expression for each $k \in [K]$ as

$$\begin{aligned} b_k(\mathbf{e}(x) + \boldsymbol{\delta}(x)) &= \frac{A_k(x)}{K + S_{\text{init}}(x)} = \frac{K + S_{\text{ret}}(x)}{q(x)(K + S_{\text{init}}(x))} \frac{q(x)A_k(x)}{K + S_{\text{ret}}(x)} \\ &= \frac{K + S_{\text{ret}}(x)}{q(x)(K + S_{\text{init}}(x))} b_k(\mathbf{e}(x)). \end{aligned} \tag{24}$$

This factorization also holds when $A_k(x) = 0$. The common coefficient c is therefore determined entirely by K, $q(x)$, and $S_{\text{init}}(x)$, rather than introduced as a model parameter. Using $S_{\text{ret}}(x) = q(x)S_{\text{init}}(x)$ gives

$$c = \frac{K + S_{\text{ret}}(x)}{q(x)(K + S_{\text{init}}(x))} = 1 + \frac{K(1 - q(x))}{q(x)(K + S_{\text{init}}(x))} > 1. \tag{25}$$

It follows that $\mathbf{b}(\mathbf{e}(x) + \boldsymbol{\delta}(x)) = c\,\mathbf{b}(\mathbf{e}(x))$. Multiplying all belief masses by c leaves each pairwise balance term and the normalized weights of competing classes unchanged. By the dissonance definition in Eq. 1,

$$\begin{aligned} D(\mathbf{b}(\mathbf{e}(x) + \boldsymbol{\delta}(x))) &= cD(\mathbf{b}(\mathbf{e}(x))) \\ &\geq D(\mathbf{b}(\mathbf{e}(x))). \end{aligned} \tag{26}$$

If at least two classes $i \neq j$ have positive initial support, then $b_i(\mathbf{e}(x))$ and $b_j(\mathbf{e}(x))$ are positive. Their pairwise balance is positive, so $D(\mathbf{b}(\mathbf{e}(x))) > 0$. Since $c > 1$, the dissonance inequality is then strict. If only one class has positive support, both opinions have zero dissonance under the zero-denominator convention. This proves both claims. □

B.3 Full Proof of Theorem 2

Theorem 2 (AS-UCE provides a gradient signal for evidence rejection)**.** *For a prompt x with $S_{\text{init}}(x) > 0$, hold the semantic frame, initial support, and augmented target $\mathbf{y}^+(x)$ fixed. The AS-UCE loss satisfies*

$$\frac{\partial \mathcal{L}_{\text{AS-UCE}}(x)}{\partial q(x)} = -\sum_{k=1}^{K} y_k^+(x) A_k(x) \psi'\big(\alpha_k^+(x)\big) + y_\perp^+(x) S_{\text{init}}(x) \psi'\big(\alpha_\perp^+(x)\big), \tag{11}$$

where ψ' is the trigamma function. The second term provides an evidence-rejection signal that is strictly positive whenever $y_\perp^+(x) > 0$. When it exceeds the magnitude of the semantic-retention term, decreasing $q(x)$ is a descent direction, reducing retained semantic evidence and increasing vacuity. In particular, $y_\perp^+(x) = 1$ guarantees $\frac{\partial \mathcal{L}_{\text{AS-UCE}}(x)}{\partial q(x)} > 0$.

Proof. We hold the semantic frame, initial support, and augmented soft reliability target $\mathbf{y}^+(x)$ fixed when differentiating with respect to $q(x)$. Recall the AS-UCE loss

$$\mathcal{L}_{\text{AS-UCE}}(x) = \sum_{k=1}^{K} y_k^+(x) \left[\psi(\alpha_0^+(x)) - \psi(\alpha_k^+(x))\right] + y_\perp^+(x) \left[\psi(\alpha_0^+(x)) - \psi(\alpha_\perp^+(x))\right]. \tag{27}$$

The existing evidence decomposition gives $e_\perp(x) = \sum_{k=1}^{K} \delta_k(x) = S_{\text{init}}(x) - S_{\text{ret}}(x)$. Thus, the augmented total strength is $\alpha_0^+(x) = K + 1 + S_{\text{init}}(x)$, independent of $q(x)$. Using the Dirichlet parameters from Section 4.3,

$$\frac{\partial \alpha_0^+(x)}{\partial q(x)} = 0, \qquad \frac{\partial \alpha_k^+(x)}{\partial q(x)} = A_k(x), \qquad \frac{\partial \alpha_\perp^+(x)}{\partial q(x)} = -S_{\text{init}}(x). \tag{28}$$

Since the target is fixed, differentiating the loss yields

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{AS-UCE}}(x)}{\partial q(x)} &= -\sum_{k=1}^{K} y_k^+(x) \psi'(\alpha_k^+(x)) \frac{\partial \alpha_k^+(x)}{\partial q(x)} - y_\perp^+(x) \psi'(\alpha_\perp^+(x)) \frac{\partial \alpha_\perp^+(x)}{\partial q(x)} \\ &= -\sum_{k=1}^{K} y_k^+(x) A_k(x) \psi'(\alpha_k^+(x)) + y_\perp^+(x) S_{\text{init}}(x) \psi'(\alpha_\perp^+(x)). \end{aligned} \tag{29}$$

The semantic-retention term is nonpositive. Because $S_{\text{init}}(x) > 0$ and $\psi'(z) > 0$ for $z > 0$, the evidence-rejection term is strictly positive whenever $y_\perp^+(x) > 0$. When it exceeds the magnitude of the retention term, the total derivative is positive. If $y_\perp^+(x) = 1$, normalization of $\mathbf{y}^+(x)$ implies $y_k^+(x) = 0$ for every semantic class, giving

$$\frac{\partial \mathcal{L}_{\text{AS-UCE}}(x)}{\partial q(x)} = S_{\text{init}}(x) \psi'(\alpha_\perp^+(x)) > 0. \tag{30}$$

In either case, decreasing $q(x)$ is a descent direction and reduces $S_{\text{ret}}(x)$. Vacuity consequently increases, since

$$\frac{\partial u(x)}{\partial q(x)} = -\frac{K S_{\text{init}}(x)}{(K + q(x) S_{\text{init}}(x))^2} < 0. \tag{31}$$

The semantic belief masses share the factor $\frac{q(x)}{K + q(x) S_{\text{init}}(x)}$, which is strictly increasing in $q(x)$. By the positive homogeneity of dissonance established in the proof of Theorem 1, lowering $q(x)$ therefore decreases dissonance whenever it is positive and leaves zero dissonance unchanged. □

B.4 Representing and Supervising Rejected Evidence

Theorem 3 (Information loss under filtered semantic UCE)**.** *Fix a K-class semantic frame. Consider a filtered construction that forms its semantic subjective opinion solely from retained evidence $\mathbf{e}(x) \in \mathbb{R}^K_{\geq 0}$ via $\boldsymbol{\alpha}(x) = \mathbf{e}(x) + \mathbf{1}_K$, without recording the rejected evidence $e_\perp(x) = \sum_{k=1}^{K} \delta_k(x)$. If $\mathbf{e}(x) = \mathbf{e}(x')$ but $e_\perp(x) \neq e_\perp(x')$, then*

$$\mathbf{b}(x) = \mathbf{b}(x'), \qquad u(x) = u(x'), \qquad D(\mathbf{b}(x)) = D(\mathbf{b}(x')). \tag{32}$$

The rejected amount cannot in general be recovered from the retained subjective opinion alone. If the UCE expression is evaluated with fixed zero semantic target weights for a fully unsupported example, then $\mathcal{L}_{\mathrm{UCE}}(x) = 0$ *and* $\nabla_\phi \mathcal{L}_{\mathrm{UCE}}(x) = \mathbf{0}$.

Proof. For a fixed semantic frame and prior, identical retained evidence yields identical Dirichlet parameters, belief masses, and vacuity. Dissonance is also identical because it is a function of the belief vector. Since the same retained opinion can correspond to different rejected amounts, no function of that opinion alone can recover $e_\perp$ for all such inputs.

For the zero-target case, substituting fixed zero semantic weights into the UCE expression gives

$$\mathcal{L}_{\mathrm{UCE}}(x) = \sum_{k=1}^{K} 0\,[\psi(\alpha_0(x)) - \psi(\alpha_k(x))] = 0 \tag{33}$$

for every ϕ, and therefore its gradient is zero. This is an algebraic extension of UCE, since the zero vector is not a normalized categorical target. □

Filtering can keep unreliable support out of the semantic frame, but the retained opinion alone does not record how much was removed. AS-UCE makes this rejected amount explicit during training and assigns fully unsupported examples a valid augmented target with $y_\perp^+ = 1$, yielding the evidence-rejection gradient in Theorem 2. Under the zero-target scenario in Theorem 3, fully unsupported examples provide no UCE learning signal. AS-UCE includes the unsupported class in its augmented soft reliability target $\mathbf{y}^+$, allowing these examples to contribute evidence-rejection supervision and making fuller use of the training data.

B.5 Additional Analysis: Dissonance Depends on Retained Belief Mass

We next show why dissonance is better aligned with supported ambiguity than entropy over normalized semantic probabilities.

Proposition 1 (Dissonance scales with retained belief mass)**.** *Let* $\mathbf{b} \in \mathbb{R}^K_{\geq 0}$ *be a belief vector and let* $c > 0$ *be a constant. Then* $D(c\mathbf{b}) = cD(\mathbf{b})$*. Therefore, if the relative belief allocation is fixed, dissonance decreases linearly as the total retained belief mass decreases.*

Proof. For $c > 0$ and $b_i + b_j > 0$,

$$\mathrm{Bal}(cb_j, cb_i) = 1 - \frac{|cb_j - cb_i|}{cb_j + cb_i} = 1 - \frac{|b_j - b_i|}{b_j + b_i} = \mathrm{Bal}(b_j, b_i). \tag{34}$$

The zero-denominator cases follow the convention in Eq. 1. For terms with positive denominators, cancellation of c gives

$$D(c\mathbf{b}) = \sum_{i=1}^{K} cb_i \frac{\sum_{j\neq i} cb_j \mathrm{Bal}(b_j, b_i)}{\sum_{j\neq i} cb_j} = cD(\mathbf{b}). \tag{35}$$

□

Entropy of a normalized semantic distribution depends only on relative allocation across semantic classes. In contrast, dissonance depends on retained belief mass. In other words, if the model has little retained support, dissonance becomes small even when normalized semantic probabilities appear diffuse.

B.6 Additional Analysis: Same Semantic Allocation Can Yield Different Decompositions

A noticeable advantage of the evidential formulation is that two prompts with similar semantic dispersion can still yield different decompositions if their retained evidence differs.

Proposition 2 (Same semantic allocation can yield different decompositions)**.** *Consider two prompts* x_a *and* x_e *with the same number* K *of semantic classes and equal total initial support,* $S_{\mathrm{init}}(x_a) = S_{\mathrm{init}}(x_e) > 0$*, and the same normalized semantic allocation:*

$$\frac{A_k(x_a)}{S_{\mathrm{init}}(x_a)} = \frac{A_k(x_e)}{S_{\mathrm{init}}(x_e)}, \quad \forall k \in [K]. \tag{36}$$

If $q(x_a) > q(x_e)$, then $u(x_a) < u(x_e)$. Moreover, if the shared allocation induces nonzero dissonance, then $D(\mathbf{b}(x_a)) > D(\mathbf{b}(x_e))$.

Proof. Since $S_{\text{init}}(x_a) = S_{\text{init}}(x_e)$ and $q(x_a) > q(x_e)$,

$$S_{\text{ret}}(x_a) = q(x_a)S_{\text{init}}(x_a) > q(x_e)S_{\text{init}}(x_e) = S_{\text{ret}}(x_e). \tag{37}$$

Since $u(x) = \frac{K}{K+S_{\text{ret}}(x)} = \frac{K}{K+q(x)S_{\text{init}}(x)}$, greater retained evidence implies lower vacuity. Hence $u(x_a) < u(x_e)$.

The two prompts have the same normalized semantic allocation, so their belief vectors differ only by the retained belief scale. Since dissonance scales linearly with retained belief mass by Proposition 1, the prompt with larger retained belief mass has larger dissonance whenever the shared allocation induces nonzero dissonance. Therefore, we have $D(\mathbf{b}(x_a)) > D(\mathbf{b}(x_e))$. □

This proposition distinguishes supported ambiguity from unsupported diversity. A genuinely ambiguous prompt can have several supported semantic alternatives and therefore high dissonance with low vacuity. A weakly supported or OOD prompt can have similar normalized semantic dispersion, but low retained evidence leads to high vacuity and lower retained dissonance.

C Pseudo-OOD Regularization

Following prior evidential deep learning work [29, 38, 41, 46], we can encourage the evidential probe to recognize insufficient knowledge by introducing pseudo-OOD samples from a different domain than the held-out OOD samples. We penalize the KL divergence between the retained semantic Dirichlet opinion and a flat Dirichlet prior:

$$\mathcal{L}_{\text{POOD}} = \text{KL}\big(\text{Dir}(\mathbf{p}|\boldsymbol{\alpha}^+_{\text{ret}}(x)) \parallel \text{Dir}(\mathbf{p} \mid \mathbf{1})\big), \tag{38}$$

where $\boldsymbol{\alpha}^+_{\text{ret}}(x) = [\alpha^+_1(x), \ldots, \alpha^+_K(x)]$ denotes the strength vector derived from retained evidence, excluding the unsupported component $\alpha^+_\perp$. We combine this regularizer with AS-UCE to train the evidential probe. The overall objective is

$$\mathcal{L}_{\text{Reg-AS-UCE}} = \frac{1}{|\mathcal{D}_{\text{IND}}|}\sum_{i=1}^{|\mathcal{D}_{\text{IND}}|}\mathcal{L}^{(i)}_{\text{AS-UCE}} + \lambda\frac{1}{|\mathcal{D}_{\text{POOD}}|}\sum_{j=1}^{|\mathcal{D}_{\text{POOD}}|}\mathcal{L}^{(j)}_{\text{POOD}}, \tag{39}$$

where $\lambda > 0$ is the trade-off hyperparameter, $\mathcal{D}_{\text{IND}}$ denotes the in-distribution training set, and $\mathcal{D}_{\text{POOD}}$ denotes the pseudo-OOD training set.

D Experiment Details

D.1 Dataset Construction

The OOD samples are drawn from designated datasets that contain entities or events outside the knowledge cutoff of the LLM backbone, while still having factual ground-truth answers. The binary OOD label is determined by whether a sample comes from an OOD dataset. In the main setting, the held-out OOD samples appear only in the test split, where they are mixed with questions from the in-distribution (IND) source. The training and validation splits contain only IND samples.

To instantiate this setting, we use `TriviaQA` [17], a short-form factual question-answering dataset with ground-truth answer annotations, as the IND dataset. For OOD datasets, we use `Alcuna` [45], which contains attributive questions about synthetic novel animals. In addition, we construct `TemporalOOD` by prompting GPT-5.4 to collect 2,000 factual questions about news and events after November 2025, paired with ground-truth answer annotations. The annotation prompt template is shown as follows:

Prompt for `TemporalOOD` annotation

You are a synthetic dataset generator. Please follow the style of TriviaQA and NQ-Open to generate 2,000 unique short-form open-domain factual question answering pairs with knowledge newer than November, 2025. You need to find unseen events and news after the knowledge cutoff to make sure pretrained LLMs have no prior knowledge of them. The questions can ask about a date, a year, a geological location, a celebrity, an entity, or an event that can be found from commonsense knowledge and is clearly defined and answerable. The responses should be designed to not have any binary answers. Do not generate rephrased questions with identical factual answers. The dataset should be returned in JSON format, with attributes of question_id (integer), question (string), and reference_answer (list of strings showing candidate answers regarding the expression variation).

The OOD label is defined as $y_i^{\text{ood}} = \mathbb{I}(x_i \in \mathcal{D}_{\text{OOD}})$, where $\mathcal{D}_{\text{OOD}}$ denotes the OOD evaluation source, either `Alcuna` or `TemporalOOD`. Here, $y_i^{\text{ood}} = 1$ indicates an OOD question and $y_i^{\text{ood}} = 0$ indicates IND. We use predicted vacuity $U_{\text{epi}}(x)$ as the uncertainty score for this task.

D.2 Descriptions of Baselines

Input Clarification Ensembling (ICE) [12] decomposes uncertainty using a two-level ensembling procedure over inputs and outputs. At the outer level, a clarification model generates multiple plausible clarifications that append missing context to the original question to reduce input ambiguity. At the inner level, the language model samples multiple responses for each clarified question. ICE interprets disagreement across clarified inputs as uncertainty attributable to ambiguity in the question, while uncertainty within each clarified question is treated as epistemic. Total uncertainty is computed from the semantic entropy of the pooled responses across all clarifications. Epistemic uncertainty is estimated by subtracting estimated aleatoric uncertainty from the total uncertainty.

Ensemble-of-Ensembles (EoE) [28] also follows a nested ensembling strategy, but instead of using explicit clarification, it constructs multiple semantically perturbed variants of the original question. For each perturbed input, it samples multiple responses and computes uncertainty in the embedding space of generated answers. Using the law of total variance, EoE treats variability among generations for the same perturbed input as aleatoric uncertainty and variability across perturbed inputs as epistemic uncertainty.

Semantic Entropy (SE) [20] estimates uncertainty over semantic meanings rather than expression variations. For each question, it samples multiple responses from the language model and groups semantically equivalent responses into prompt-specific semantic clusters using a semantic clustering routine $\texttt{Cluster}(\cdot)$, typically using bidirectional entailment in natural language inference (NLI) or semantic equivalence judgments. It then forms an empirical distribution $\hat{\mathbf{p}}^{\text{sem}}(x)$ over the observed clusters and computes

$$H_{\text{sem}}(x) = -\sum_{k=1}^{K} \hat{p}_k^{\text{sem}}(x) \log \hat{p}_k^{\text{sem}}(x), \tag{40}$$

where K is the number of semantic clusters for prompt x. A higher semantic entropy indicates that the model spreads probability mass across multiple distinct meanings. In our experiments, we use this scalar score for both OOD detection and ambiguity detection.

SelfCheckGPT [31] is a black-box consistency-based method for detecting unreliable generations with an uncertainty score. Given a question and a candidate response, it samples additional responses from the same language model and measures whether the candidate response is supported by the sampled responses. The support score can be computed using entailment-based comparison. We then use the average inconsistency between the candidate response and the sampled responses as a prompt-level uncertainty score, where higher values indicate less stable model behavior across repeated generations.

Kernel Language Entropy (KLE) [35] extends semantic entropy by incorporating fine-grained semantic relationships between generated responses. Given multiple sampled responses, KLE constructs a positive semi-definite and unit-trace semantic kernel matrix that encodes pairwise semantic similarities, derived from an NLI-based semantic graph. It quantifies uncertainty using the

von Neumann entropy of the graph kernel. Unlike semantic entropy, which treats distinct semantic clusters as separate categories, KLE accounts for soft-assignment of response similarities.

EigenValue (EigV) [25] is another sampling-based approach that estimates uncertainty from the spectral properties of a semantic similarity graph. It samples multiple responses and constructs a weighted graph whose nodes represent generated responses and whose edge weights encode pairwise semantic similarities. EigV computes the eigenvalues of the normalized graph Laplacian and aggregates them as $U_{\text{EigV}} = \sum_{i=1}^{M} \max(0, 1 - \lambda_i)$, where λ_i denotes the i-th eigenvalue and M is the number of sampled responses. The resulting score provides a continuous approximation to the number of distinct semantic groups, with higher values indicating greater semantic dispersion.

Semantic Density (SD) [37] estimates response-level confidence using an approach inspired by kernel density estimation in semantic space. Given a target response, SD obtains multiple diverse reference responses using diverse beam search and estimates their semantic relationships with the target using NLI. It then estimates the target response's semantic density by aggregating a modified Epanechnikov kernel evaluated at these distances, weighted by the length-normalized generation probabilities of the reference responses. A higher semantic density indicates that the target response is supported by semantically similar and highly probable alternatives, and therefore less uncertain.

CoCoA [42] combines model confidence and semantic consistency within a minimum Bayes risk framework. Given an input x and a target response, it samples multiple responses from the language model and measures their semantic similarity to the target response. CoCoA then computes a response-level uncertainty score by multiplying a likelihood-based uncertainty measure, for example, negative log-likelihood, by the average semantic disagreement between the target response and the sampled responses. Higher uncertainty is assigned to responses that have low model confidence and weak semantic agreement with alternative generations.

D.3 Implementation Details

For each question, we sample $M = 10$ responses and map them to prompt-specific semantic classes using the clustering pipeline from the semantic entropy implementation [20]. Each pair of responses is checked with the NLI model `DeBERTa-v2-xlarge-MNLI` in both directions. Two responses count as semantically equivalent if neither direction is predicted as contradiction and at least one direction is predicted as entailment. Classes are then formed greedily in sampling order. A response that has not been assigned yet opens a new class, and every later response equivalent to it joins that class. To construct the soft semantic reliability target $\mathbf{y}^+$, the reliability score of a prompt can be taken directly from correctness judgments or predicted by an auxiliary probe trained on them. Our experiments use the auxiliary probe, trained on correctness judgments from `gpt-5-mini`. We implement it as a two-layer MLP, although any probe architecture can serve this role. The auxiliary probe and the evidential probe both take the hidden states of the last five layers at the token immediately before generation as input. We select the discount strength from the grid $\zeta \in \{0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 5.0, 10.0\}$ by maximizing the validation AUROC of the task-relevant uncertainty component. When pseudo-OOD samples are unavailable for OOD detection, we instead select ζ by minimizing the validation loss. We train the evidential probe, parameterized as logistic regression, for 400 epochs using the Adam optimizer with a learning rate of 5×10^{-4}. All questions are prompted with a 0-shot template. The experiments were conducted on 8 NVIDIA A6000 GPUs.

The non-regularized setting uses $4{,}000/500/500$ questions for training, validation, and testing, respectively. Pseudo-OOD regularization adds $1,600/200$ pseudo-OOD samples to the training and validation splits, respectively. For the regularizer trade-off coefficient, we sweep $\lambda \in \{0.001, 0.01, 0.1, 0.25, 1.0\}$ and use a single shared value $\lambda = 0.25$ for all backbones and datasets. To study the effect of pseudo-OOD regularization, we fix ζ to the best value selected for each backbone–dataset pair at $\lambda = 0.25$.

For the uncertainty decomposition baselines, we retain most of the original ICE settings but change the clarification generator to `gpt-5-mini`. To reduce computational overhead for EoE, we reduce the budget to 5 perturbations and 10 responses per perturbed input, while retaining `sentence-transformers/all-MiniLM-L6-v2` as the embedding model and `eugenesiow/bart-paraphrase` as the paraphrase model.

E Efficient Evidential Probing with Iterative Sampling

In settings with a small sampling budget, standard repeated sampling may produce semantically redundant responses and fail to expose distinct candidate meanings. To improve semantic coverage, we introduce EPN-IT, an inference-time iterative sampling strategy that explicitly encourages the backbone LLM to generate responses with different semantic meanings. The first response is generated with the original prompt template, so that it reflects the model's default answering behavior. Each subsequent response is conditioned on the previously generated answers and is prompted to express a genuinely different semantic meaning.

We additionally allow the model to output `NONE` when it cannot produce a new semantic meaning. This refusal option allows the model to decline an additional answer, reducing the pressure to manufacture artificial diversity. Responses equal to `NONE` are not treated as semantic classes and are excluded from evidence estimation. Thus, iterative sampling expands the prompt-specific semantic frame only through accepted non-`NONE` responses.

After iterative generation, we treat each accepted non-`NONE` response as a proxy semantic class, without an additional clustering step, and apply EPN's evidence-estimation pipeline to form a subjective opinion.

The EPN-IT prompt templates are shown below. The first generated response uses the original prompt template:

Prompt for iterative sampling's primary response

{question}

Respond with the answer only. Do not explain.

For the remaining responses, we condition on the previously generated answers and encourage semantic distinctiveness:

Prompt for the k-th response in iterative sampling

{question}

Here are your previous responses:
{former sampled response 1}
...
{former sampled response $k-1$}

Generate one new response whose underlying semantic meaning or interpretation is genuinely different from all previous responses. Do not paraphrase, restate, or slightly reword an existing response. A different response must express a different interpretation, hypothesis, entity, value, or conclusion. If you cannot produce a genuinely different meaning, output **NONE**.

Respond with the answer only. Do not explain.

F Case Study on Ambiguous Questions

In this section, we present examples from `AmbigQA` to examine the model's response behavior. We first show an example in which the question is annotated as ambiguous, but the model's behavior is dominated by a lack of knowledge. This creates an uncertainty overriding pattern in which epistemic uncertainty masks the ambiguity signal, leading to a mismatch between dissonance and the ambiguity label. The predicted vacuity and dissonance shown here are from EPN trained on Llama-3-8B-Instruct responses.

Case study: ambiguous question with ambiguity masked by insufficient knowledge

Question. Who played the baby in Baby's Day Out?

Ambiguity label. Ambiguous

References.

- Adam Robert and Jacob Joseph Worton
- Verne Troyer
- Verne Jay Troyer
- Adam Robert
- Jacob Joseph Worton

Sampled responses in semantic clusters.

1. **Cluster 1** (1 response): Blake and Dylan Tuomy-Wilhoit
2. **Cluster 2** (2 responses): Benjamin King; Benjamin King
3. **Cluster 3** (1 response): Tiger Allen
4. **Cluster 4** (1 response): Matthew Lawrence
5. **Cluster 5** (1 response): Benjamin Barker
6. **Cluster 6** (1 response): Bryan Madorsky
7. **Cluster 7** (1 response): Lauryn Alisa McClain
8. **Cluster 8** (1 response): Chris Farley
9. **Cluster 9** (1 response): Liliana Mumy

Predicted vacuity: 0.7316
Predicted dissonance: 0.2006

The question is annotated as ambiguous because "played the baby" can refer to different forms of portrayal, as reflected by the multiple reference answers. However, the sampled responses do not concentrate on a small set of plausible interpretations. Instead, they scatter across many unrelated semantic clusters, most of which are unsupported by the references. This suggests that the model is not expressing uncertainty over competing valid interpretations, but is instead guessing due to insufficient knowledge. In such cases, epistemic uncertainty can override the ambiguity signal, and the retained evidence for each semantic alternative becomes weak, increasing vacuity and reducing the conflict required for high dissonance. This explains why annotated ambiguous questions can be difficult to detect, especially for smaller models whose response behavior may be dominated by lack of knowledge rather than ambiguity among plausible meanings.

In contrast, the following example exhibits high dissonance when the question is ambiguous and the model has the knowledge to answer it.

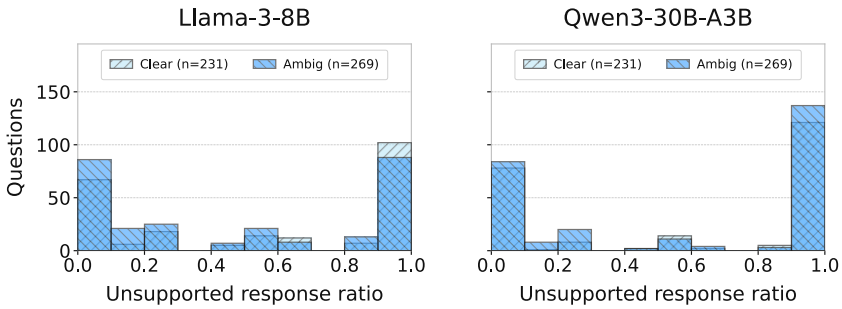


Figure 3: Histogram of per-question unsupported response ratios on `AmbigQA` for Llama-3-8B and Qwen3-30B-A3B test generations.

Case study: ambiguous question with well-supported semantic alternatives

Question. Vikram Samvat is official calendar of which contry?

Ambiguity label. Ambiguous

References.

- Nepal
- Federal Democratic Republic of Nepal
- Republic of India
- India

Sampled responses in semantic clusters.

1. **Cluster 1** (5 responses): Nepal; Nepal; Nepal; Nepal; Nepal
2. **Cluster 2** (5 responses): India; India; India; India; India

Predicted vacuity: 0.2646
Predicted dissonance: 0.5102

This example illustrates the intended high-dissonance behavior for ambiguity detection. The question is ambiguous because `Vikram Samvat` can be associated with more than one valid country-level answer, as reflected by the references containing both Nepal and India. Unlike the previous case, the sampled responses do not scatter across unrelated entities. Instead, they concentrate on two semantically distinct and reference-supported clusters, with balanced support for Nepal and India. This indicates that the model has sufficient knowledge of the plausible interpretations, but the prompt itself does not determine which interpretation should be selected. As a result, retained semantic evidence remains high for multiple alternatives, leading to low vacuity and high dissonance. This case supports the role of dissonance as an ambiguity signal: it becomes informative when semantic diversity arises from competing plausible meanings rather than from unsupported guessing.

To further understand the low absolute AUROC of dissonance for ambiguity detection, we compute the per-question unsupported response ratio as the fraction of sampled responses that are unsupported. Let $\mathcal{G}(x) = \{\texttt{ref}_1(x), \ldots, \texttt{ref}_L(x)\}$ be the set of L ground-truth references for input x. The ratio is

$$\mathrm{ur}(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\{\max_j \texttt{RougeL}(r^{(m)}, \texttt{ref}_j(x)) \leq 0.3\}. \tag{41}$$

Here, we define a reference-supported response as one whose ROUGE-L score against at least one reference is > 0.3. A high value indicates that many sampled responses are not valid interpretations of the question, even if the question is annotated as ambiguous. Figure 3 shows that, for both the 8B and 30B backbones, nearly half of the ambiguous questions have high unsupported response ratios, indicating that the model often responds to annotated ambiguous questions with invalid guesses rather than a small set of reference-supported alternatives. This pattern helps explain why dissonance can be weak for ambiguity detection. When many sampled responses are unsupported, AS-UCE routes their initial support away from the semantic frame, which increases vacuity and weakens the

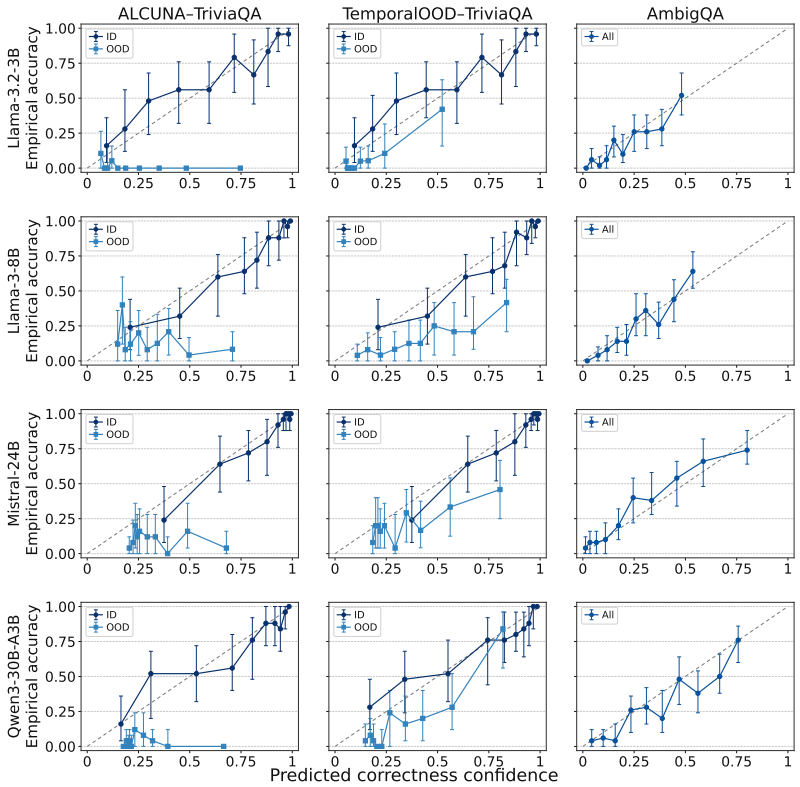


Figure 4: Reliability diagrams of the auxiliary-probe correctness confidence. Each curve uses ten equal-frequency bins and shows empirical accuracy with 95% bootstrap intervals. The dashed diagonal indicates perfect calibration.

retained semantic conflict. Thus, some ambiguous questions exhibit an uncertainty-overriding pattern: epistemic uncertainty dominates the response behavior and masks the ambiguity signal. This does not contradict the intended role of dissonance; rather, it shows that dissonance is most informative when semantic diversity comes from supported competing meanings, and less effective when annotated ambiguity is mixed with lack of model knowledge.

G Additional Experiments

G.1 Beyond OOD Detection with Calibration and Selective Prediction

Here we assess the calibration of the correctness confidence from EPN's auxiliary probe and examine error detection and selective answering beyond OOD and ambiguity detection. This confidence is produced by the auxiliary probe trained on correctness judgments. The backbone likelihood plays a different role, since the length-normalized likelihood $s^{(m)}_{\mathrm{norm}}(x)$ only serves as initial support and the evidential probe decides how much of it to retain. We use EPN trained with AS-UCE without pseudo-OOD regularization on seed 42. On the two OOD benchmarks, we exclude prompts whose greedy answer is a refusal, because calibration compares the probe confidence with the correctness of a committed answer, and a refusal offers no such answer. This exclusion applies to this section only, so the OOD accuracies here are lower than in Table 19, which counts all prompts over five seeds. For ALCUNA–TriviaQA and TemporalOOD–TriviaQA, we plot the TriviaQA (ID) and OOD subsets separately, so source discrimination alone cannot account for the within-source calibration results. Correctness labels come from the same GPT-5-mini judge used for training, and error bars are 95% bootstrap intervals.

Confidence calibration. Figure 4 plots answer accuracy against the probe confidence, and Table 5 lists the ECE and Brier score. On the ID subsets, ECE stays between 0.038 and 0.082 for all four

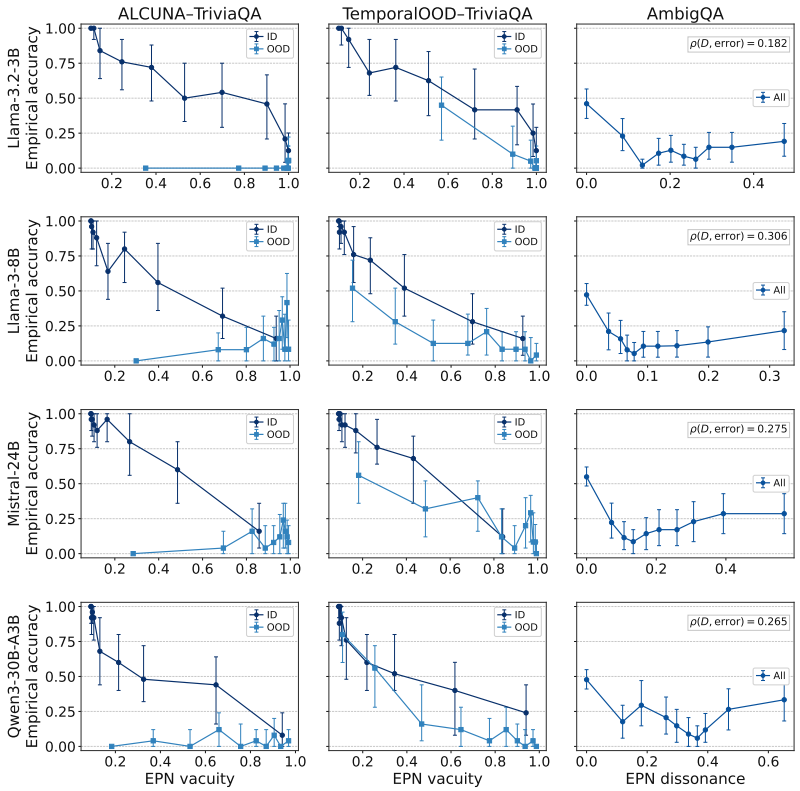


Figure 5: Answer accuracy against the task-relevant EPN uncertainty: vacuity for the two OOD benchmarks (ID and OOD subsets) and dissonance for AmbigQA. The OOD benchmarks use ten equal-frequency bins; AmbigQA puts all zero-dissonance prompts in one bin and splits the rest into nine equal-frequency bins. Error bars are 95% bootstrap intervals, and $\rho(D, \text{error})$ is the Spearman correlation between dissonance and answer error.

backbones, and on AmbigQA between 0.050 and 0.076, so the probe confidence closely tracks how often the answers are correct. The OOD subsets tell a different story, with ECE ranging from 0.080 to 0.255. The probe already gives low confidence to most OOD prompts, yet accuracy falls even lower, which leaves the curves below the diagonal. This is the regime that vacuity is designed to identify.

Accuracy against EPN uncertainty. Figure 5 examines whether prompts with higher EPN task-relevant uncertainty are answered incorrectly more often. We treat vacuity and dissonance as rankings here, not as probabilities. On the ID subsets, accuracy falls steadily as vacuity grows, with Spearman correlations between vacuity and error from 0.523 to 0.606. On TemporalOOD, the OOD curves sit below the ID curves but show the same positive association between vacuity and error, with correlations of 0.335, 0.335, 0.320, and 0.530 for Llama-3.2-3B, Llama-3-8B, Mistral-Small-24B, and Qwen3-30B-A3B, respectively. On AmbigQA, higher dissonance is associated with more errors for every backbone ($\rho(D, \text{error})$ of 0.182, 0.306, 0.275, and 0.265). Prompts with zero dissonance are the most accurate (0.461 to 0.549); accuracy is lower in the positive-dissonance bins, though it does not keep falling at the largest values. Dissonance therefore gives a useful error ranking on AmbigQA. ALCUNA is the one exception to the positive association between vacuity and error. The backbones answer only 1.6% to 14.6% of its OOD prompts correctly while vacuity is concentrated near one, and the correlation between vacuity and error within this subset turns slightly negative (-0.179 to -0.023). Detecting OOD prompts and ranking errors within a low-accuracy source are thus distinct tasks, and on ALCUNA vacuity serves the first. Appendix G.10 reports the same pattern across five seeds.

Selective prediction. Figure 6 turns the same rankings into an abstention experiment by answering prompts in order of increasing uncertainty up to a specified coverage threshold. We define the selective risk as the error rate among the answered prompts, and AURC averages it over all coverage levels, so lower is better and the dashed line marks the error rate without abstention. On the ID

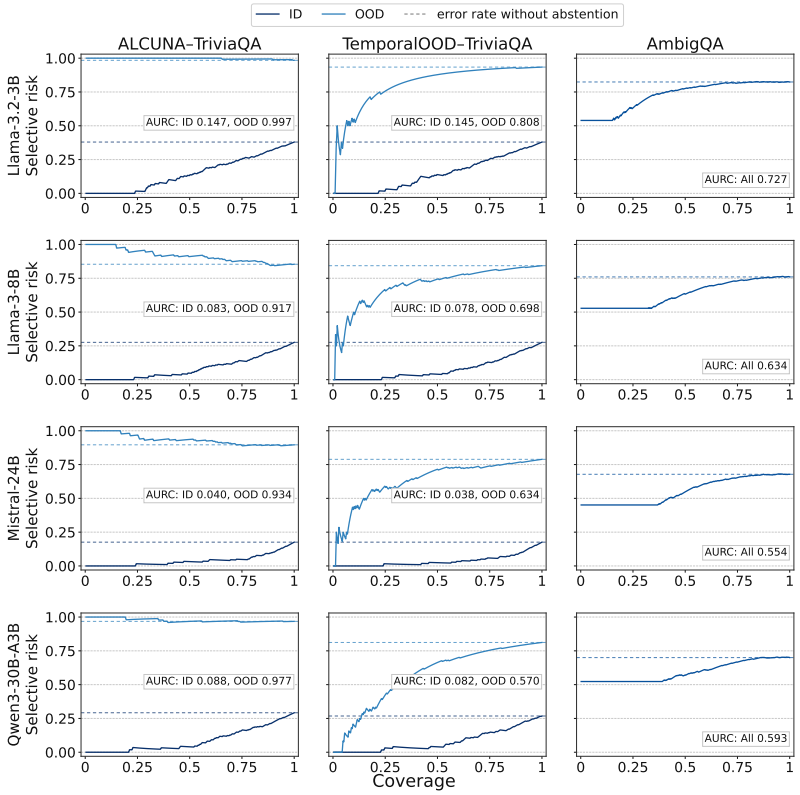


Figure 6: Risk–coverage curves. Prompts are answered from lowest to highest task-relevant uncertainty (vacuity for the OOD benchmarks, dissonance for AmbigQA). Coverage is the fraction of prompts answered and selective risk is the error rate among them; dashed lines mark each subset's error rate without any abstention. Tied scores are averaged over random orderings, which produces the flat start on AmbigQA. AURC is the mean selective risk over coverage.

datasets, the first quarter of prompts is answered almost without error, and AURC ranges from 0.038 to 0.147 against error rates of 0.176 to 0.380. On TemporalOOD, risk rises smoothly with coverage, and AURC (0.570 to 0.808) stays below the error rate (0.789 to 0.934). On AmbigQA, answering the zero-dissonance prompts first gives a risk of 0.451 to 0.539 against overall error rates of 0.678 to 0.824, with AURC between 0.554 and 0.727. On ALCUNA, the curves stay close to or above the reference at every coverage level, matching the negative correlations above. Taken together, vacuity and dissonance support error detection and risk-aware abstention on the ID, TemporalOOD, and AmbigQA subsets. ALCUNA marks the boundary of this use, since vacuity saturates near one there and leaves little room to rank prompts.

Table 5: Calibration and selective prediction results from a single run. ECE and the Brier score evaluate the auxiliary-probe confidence, with ECE computed over ten equal-frequency bins. $\rho(U, \text{error})$ is the Spearman correlation between the task-relevant uncertainty and answer error, and AURC is the area under the risk–coverage curve for the same uncertainty, which is vacuity for the OOD benchmarks and dissonance for AmbigQA. n is the number of evaluated prompts.

Backbone	Benchmark (subset)	n	Accuracy	ECE	Brier	$\rho(U, \text{error})$	AURC
Llama-3.2-3B	ALCUNA–TriviaQA (ID)	245	0.620	0.081	0.180	0.583	0.147
	ALCUNA–TriviaQA (OOD)	189	0.016	0.244	0.120	-0.153	0.997
	TemporalOOD–TriviaQA (ID)	245	0.620	0.082	0.180	0.592	0.145
	TemporalOOD–TriviaQA (OOD)	197	0.066	0.080	0.055	0.335	0.808
	AmbigQA	500	0.176	0.050	0.125	0.182	0.727
Llama-3-8B	ALCUNA–TriviaQA (ID)	250	0.724	0.056	0.136	0.577	0.083
	ALCUNA–TriviaQA (OOD)	246	0.146	0.218	0.195	-0.179	0.917
	TemporalOOD–TriviaQA (ID)	250	0.724	0.063	0.135	0.606	0.078
	TemporalOOD–TriviaQA (OOD)	242	0.157	0.254	0.200	0.335	0.698
	AmbigQA	500	0.240	0.051	0.149	0.306	0.634
Mistral-Small-24B	ALCUNA–TriviaQA (ID)	250	0.824	0.038	0.094	0.523	0.040
	ALCUNA–TriviaQA (OOD)	250	0.104	0.231	0.173	-0.100	0.934
	TemporalOOD–TriviaQA (ID)	250	0.824	0.038	0.094	0.534	0.038
	TemporalOOD–TriviaQA (OOD)	246	0.211	0.134	0.179	0.320	0.634
	AmbigQA	500	0.322	0.054	0.158	0.275	0.554
Qwen3-30B-A3B	ALCUNA–TriviaQA (ID)	250	0.708	0.059	0.151	0.603	0.088
	ALCUNA–TriviaQA (OOD)	250	0.032	0.255	0.120	-0.023	0.977
	TemporalOOD–TriviaQA (ID)	250	0.732	0.064	0.151	0.562	0.082
	TemporalOOD–TriviaQA (OOD)	250	0.188	0.153	0.126	0.530	0.570
	AmbigQA	500	0.300	0.076	0.171	0.265	0.593

G.2 Uncertainty Decomposition Validity AUPR Results

Table 6: AUPR results for uncertainty decomposition validity. Each cell reports the mean ± sample standard deviation of AUPR over five data-split seeds. EPN is trained with AS-UCE, and its dissonance (D) and vacuity (V) are the aleatoric (AU) and epistemic (EU) uncertainty components. Blue and light blue cells mark the best and runner-up means in each row, where ICE, EoE, and EPN enter with EU for OOD detection and with AU for ambiguity detection.

Dataset	Backbone	SE	SelfCheck GPT	KLE	EigV	CoCoA		Semantic Density	ICE		EoE		EPN w/ AS-UCE	
						MaxProb	PPL		AU	EU	AU	EU	D (AU)	V (EU)
Alcuna-TriviaQA	Llama-3.2-3B	0.684 ±0.018	0.630 ±0.017	0.651 ±0.013	0.619 ±0.011	0.568 ±0.011	0.539 ±0.014	0.549 ±0.013	0.590 ±0.016	**0.785 ±0.019**	0.584 ±0.005	0.583 ±0.029	0.372 ±0.025	0.640 ±0.021
	Llama-3-8B	**0.838 ±0.012**	0.807 ±0.019	0.826 ±0.006	0.791 ±0.016	0.775 ±0.021	0.739 ±0.018	0.764 ±0.018	0.525 ±0.015	**0.838 ±0.014**	0.625 ±0.008	0.587 ±0.024	0.540 ±0.007	0.812 ±0.029
	Mistral-Small-24B	0.907 ±0.012	0.874 ±0.007	0.901 ±0.010	0.891 ±0.016	0.907 ±0.007	0.884 ±0.006	0.883 ±0.009	0.707 ±0.123	0.906 ±0.021	0.790 ±0.008	0.660 ±0.032	0.633 ±0.037	**0.919 ±0.026**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.582 ±0.006	0.535 ±0.023	0.557 ±0.005	0.553 ±0.007	0.552 ±0.024	0.512 ±0.024	0.522 ±0.025	0.426 ±0.003	0.664 ±0.008	0.739 ±0.009	0.637 ±0.012	0.381 ±0.006	**0.717 ±0.021**
	Llama-3-8B	0.725 ±0.008	0.726 ±0.014	0.734 ±0.008	0.746 ±0.011	0.737 ±0.014	0.709 ±0.017	0.741 ±0.016	0.600 ±0.016	0.725 ±0.008	0.767 ±0.011	0.680 ±0.026	0.544 ±0.018	**0.784 ±0.003**
	Mistral-Small-24B	0.830 ±0.011	0.816 ±0.010	0.825 ±0.009	0.854 ±0.009	0.839 ±0.005	0.806 ±0.003	0.827 ±0.010	0.588 ±0.029	0.834 ±0.016	0.874 ±0.005	0.687 ±0.025	0.599 ±0.023	**0.867 ±0.028**
AmbigQA	Llama-3.2-3B	0.547 ±0.014	0.544 ±0.015	0.532 ±0.008	0.534 ±0.013	0.533 ±0.010	0.535 ±0.017	0.529 ±0.016	0.568 ±0.016	0.574 ±0.031	0.531 ±0.029	0.567 ±0.024	**0.598 ±0.028**	0.535 ±0.013
	Llama-3-8B	0.547 ±0.013	0.561 ±0.011	0.545 ±0.012	0.550 ±0.010	0.553 ±0.012	0.556 ±0.020	0.554 ±0.015	0.517 ±0.012	0.495 ±0.015	0.532 ±0.035	0.573 ±0.024	**0.586 ±0.026**	0.556 ±0.014
	Mistral-Small-24B	0.556 ±0.026	0.564 ±0.016	0.549 ±0.019	0.551 ±0.021	0.564 ±0.021	0.565 ±0.020	0.561 ±0.024	0.515 ±0.012	0.468 ±0.007	0.559 ±0.019	0.583 ±0.018	**0.594 ±0.022**	0.549 ±0.017

Table 6 repeats the comparison of Table 2 with AUPR. EPN dissonance has the highest mean AUPR in all three AmbigQA settings, and EPN vacuity leads the task-relevant scores in four of six OOD settings. The exceptions are the two Llama backbones on ALCUNA, where SE, KLE, and the

epistemic uncertainty of ICE reach a higher AUPR. Appendix G.3 reports bootstrap confidence intervals for these AUPR comparisons.

G.3 Bootstrap Confidence Intervals for the Main Results

Tables 2 and 6 report the mean and standard deviation of AUROC and AUPR over five data-split seeds. Here, we show additional 95% bootstrap confidence intervals for the same results. Each interval comes from 2,000 stratified bootstrap replicates. Every replicate first resamples the five seeds with replacement and then resamples the positive and negative test examples within each selected seed, so the interval reflects both the split-to-split and the within-split sampling variation. Tables 7 and 8 give the AUROC intervals, and Tables 9 and 10 give the AUPR intervals. The intervals support the conclusions of Section 6. EPN vacuity has the highest mean AUROC in all six OOD settings, and EPN dissonance has the highest mean AUROC on AmbigQA for all three backbones. On TemporalOOD, the EPN interval lies entirely above that of the strongest baseline. On ALCUNA and AmbigQA the margins are smaller and the intervals overlap, yet EPN ranks first in all nine settings, so its advantage is consistent.

Table 7: AUROC with 95% bootstrap confidence intervals for SE, SelfCheckGPT, and the decomposition methods. Each cell reports the mean ± sample standard deviation over five data-split seeds, with the interval in brackets. EPN is trained with AS-UCE, and its dissonance (D) and vacuity (V) are the AU and EU components. Blue and light blue cells mark the best and runner-up means in each row, where ICE, EoE, and EPN enter with EU for OOD detection and with AU for ambiguity detection.

Dataset	Backbone	SE	SelfCheck GPT	ICE		EoE		EPN	
				AU	EU	AU	EU	D (AU)	V (EU)
Alcuna-TriviaQA	Llama-3.2-3B	0.796 ± 0.010 [0.775, 0.816]	0.767 ± 0.008 [0.746, 0.788]	0.637 ± 0.016 [0.612, 0.661]	0.802 ± 0.016 [0.781, 0.821]	0.652 ± 0.006 [0.630, 0.675]	0.643 ± 0.023 [0.616, 0.671]	0.502 ± 0.015 [0.475, 0.529]	**0.813 ± 0.007 [0.794, 0.833]**
	Llama-3-8B	0.864 ± 0.008 [0.847, 0.879]	0.841 ± 0.006 [0.823, 0.857]	0.535 ± 0.020 [0.507, 0.558]	0.832 ± 0.014 [0.813, 0.852]	0.710 ± 0.004 [0.689, 0.731]	0.676 ± 0.017 [0.651, 0.700]	0.654 ± 0.010 [0.631, 0.680]	**0.880 ± 0.016 [0.863, 0.900]**
	Mistral-Small-24B	0.916 ± 0.003 [0.904, 0.927]	0.901 ± 0.004 [0.888, 0.913]	0.720 ± 0.122 [0.603, 0.788]	0.902 ± 0.020 [0.880, 0.919]	0.851 ± 0.005 [0.834, 0.867]	0.771 ± 0.021 [0.745, 0.796]	0.765 ± 0.035 [0.728, 0.795]	**0.936 ± 0.016 [0.921, 0.953]**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.680 ± 0.014 [0.654, 0.704]	0.624 ± 0.019 [0.597, 0.652]	0.311 ± 0.011 [0.290, 0.331]	0.693 ± 0.010 [0.669, 0.716]	0.794 ± 0.005 [0.777, 0.811]	0.728 ± 0.009 [0.706, 0.748]	0.457 ± 0.011 [0.429, 0.484]	**0.858 ± 0.006 [0.840, 0.874]**
	Llama-3-8B	0.790 ± 0.004 [0.772, 0.808]	0.780 ± 0.004 [0.762, 0.799]	0.682 ± 0.013 [0.658, 0.706]	0.757 ± 0.012 [0.736, 0.777]	0.823 ± 0.002 [0.805, 0.839]	0.768 ± 0.020 [0.744, 0.792]	0.658 ± 0.017 [0.633, 0.684]	**0.855 ± 0.007 [0.838, 0.871]**
	Mistral-Small-24B	0.852 ± 0.002 [0.837, 0.867]	0.852 ± 0.004 [0.835, 0.867]	0.626 ± 0.034 [0.592, 0.657]	0.847 ± 0.007 [0.829, 0.862]	0.888 ± 0.003 [0.874, 0.900]	0.792 ± 0.014 [0.770, 0.814]	0.731 ± 0.025 [0.700, 0.760]	**0.897 ± 0.015 [0.880, 0.914]**
AmbigQA	Llama-3.2-3B	0.474 ± 0.009 [0.451, 0.498]	0.478 ± 0.011 [0.455, 0.502]	0.509 ± 0.025 [0.478, 0.537]	0.516 ± 0.033 [0.479, 0.548]	0.459 ± 0.024 [0.429, 0.489]	0.494 ± 0.020 [0.466, 0.521]	**0.535 ± 0.025 [0.505, 0.565]**	0.469 ± 0.010 [0.444, 0.494]
	Llama-3-8B	0.510 ± 0.013 [0.485, 0.536]	0.500 ± 0.013 [0.476, 0.525]	0.503 ± 0.006 [0.490, 0.517]	0.498 ± 0.020 [0.471, 0.525]	0.484 ± 0.027 [0.454, 0.514]	0.515 ± 0.030 [0.482, 0.548]	**0.530 ± 0.009 [0.508, 0.553]**	0.513 ± 0.013 [0.490, 0.539]
	Mistral-Small-24B	0.514 ± 0.025 [0.486, 0.545]	0.518 ± 0.016 [0.491, 0.544]	0.500 ± 0.017 [0.482, 0.518]	0.471 ± 0.002 [0.449, 0.493]	0.516 ± 0.016 [0.489, 0.543]	0.527 ± 0.007 [0.504, 0.551]	**0.540 ± 0.016 [0.512, 0.565]**	0.512 ± 0.018 [0.483, 0.541]

Table 8: AUROC with 95% bootstrap confidence intervals for the additional UQ baselines. Each cell reports the mean ± sample standard deviation over five data-split seeds, with the interval in brackets. EPN is trained with AS-UCE. Blue and light blue cells mark the best and runner-up means in each row, where EPN enters with EU for OOD detection and with AU for ambiguity detection.

Dataset	Backbone	KLE	EigV	CoCoA		Semantic Density	EPN	
				MaxProb	PPL		D (AU)	V (EU)
Alcuna-TriviaQA	Llama-3.2-3B	0.697 ± 0.020 [0.671, 0.724]	0.683 ± 0.018 [0.658, 0.707]	0.646 ± 0.013 [0.622, 0.670]	0.595 ± 0.015 [0.571, 0.619]	0.609 ± 0.016 [0.583, 0.633]	0.502 ± 0.015 [0.475, 0.529]	**0.813 ± 0.007 [0.794, 0.833]**
	Llama-3-8B	0.863 ± 0.007 [0.848, 0.877]	0.853 ± 0.008 [0.836, 0.869]	0.845 ± 0.010 [0.828, 0.863]	0.803 ± 0.009 [0.785, 0.821]	0.827 ± 0.009 [0.809, 0.844]	0.654 ± 0.010 [0.631, 0.680]	**0.880 ± 0.016 [0.863, 0.900]**
	Mistral-Small-24B	0.912 ± 0.004 [0.900, 0.924]	0.916 ± 0.003 [0.904, 0.927]	0.925 ± 0.002 [0.915, 0.935]	0.919 ± 0.002 [0.907, 0.930]	0.909 ± 0.002 [0.898, 0.920]	0.765 ± 0.035 [0.728, 0.795]	**0.936 ± 0.016 [0.921, 0.953]**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.580 ± 0.012 [0.556, 0.606]	0.558 ± 0.013 [0.535, 0.583]	0.587 ± 0.016 [0.561, 0.614]	0.539 ± 0.017 [0.513, 0.566]	0.546 ± 0.016 [0.519, 0.572]	0.457 ± 0.011 [0.429, 0.484]	**0.858 ± 0.006 [0.840, 0.874]**
	Llama-3-8B	0.797 ± 0.003 [0.780, 0.815]	0.800 ± 0.004 [0.781, 0.818]	0.793 ± 0.007 [0.774, 0.811]	0.755 ± 0.013 [0.732, 0.776]	0.796 ± 0.008 [0.777, 0.814]	0.658 ± 0.017 [0.633, 0.684]	**0.855 ± 0.007 [0.838, 0.871]**
	Mistral-Small-24B	0.851 ± 0.003 [0.835, 0.866]	0.857 ± 0.004 [0.841, 0.873]	0.860 ± 0.002 [0.845, 0.875]	0.847 ± 0.001 [0.830, 0.862]	0.851 ± 0.003 [0.836, 0.867]	0.731 ± 0.025 [0.700, 0.760]	**0.897 ± 0.015 [0.880, 0.914]**
AmbigQA	Llama-3.2-3B	0.459 ± 0.007 [0.435, 0.484]	0.459 ± 0.008 [0.435, 0.483]	0.464 ± 0.011 [0.439, 0.489]	0.469 ± 0.014 [0.444, 0.496]	0.457 ± 0.012 [0.433, 0.480]	**0.535 ± 0.025 [0.505, 0.565]**	0.469 ± 0.010 [0.444, 0.494]
	Llama-3-8B	0.503 ± 0.015 [0.477, 0.526]	0.512 ± 0.015 [0.485, 0.537]	0.506 ± 0.011 [0.481, 0.530]	0.506 ± 0.014 [0.482, 0.532]	0.506 ± 0.010 [0.481, 0.529]	**0.530 ± 0.009 [0.508, 0.553]**	0.513 ± 0.013 [0.490, 0.539]
	Mistral-Small-24B	0.509 ± 0.021 [0.482, 0.537]	0.515 ± 0.021 [0.487, 0.544]	0.527 ± 0.021 [0.500, 0.556]	0.521 ± 0.020 [0.494, 0.550]	0.521 ± 0.022 [0.494, 0.550]	**0.540 ± 0.016 [0.512, 0.565]**	0.512 ± 0.018 [0.483, 0.541]

Table 9: AUPR with 95% bootstrap confidence intervals for SE, SelfCheckGPT, and the decomposition methods. Each cell reports the mean ± sample standard deviation over five data-split seeds, with the interval in brackets. OOD and ambiguous prompts form the positive class. EPN is trained with AS-UCE, and its dissonance (D) and vacuity (V) are the AU and EU components. Blue and light blue cells mark the best and runner-up means in each row, where ICE, EoE, and EPN enter with EU for OOD detection and with AU for ambiguity detection.

Dataset	Backbone	SE	SelfCheck GPT	ICE		EoE		EPN	
				AU	EU	AU	EU	D (AU)	V (EU)
Alcuna-TriviaQA	Llama-3.2-3B	0.684 ± 0.018 [0.656, 0.719]	0.630 ± 0.017 [0.604, 0.669]	0.590 ± 0.016 [0.569, 0.619]	**0.785 ± 0.019 [0.762, 0.811]**	0.584 ± 0.005 [0.567, 0.613]	0.583 ± 0.029 [0.557, 0.619]	0.372 ± 0.025 [0.357, 0.404]	0.640 ± 0.021 [0.613, 0.679]
	Llama-3-8B	**0.838 ± 0.012 [0.817, 0.860]**	0.807 ± 0.019 [0.782, 0.836]	0.525 ± 0.015 [0.507, 0.552]	**0.838 ± 0.014 [0.819, 0.858]**	0.625 ± 0.008 [0.606, 0.653]	0.587 ± 0.024 [0.564, 0.619]	0.540 ± 0.007 [0.526, 0.565]	0.812 ± 0.029 [0.784, 0.851]
	Mistral-Small-24B	0.907 ± 0.012 [0.889, 0.924]	0.874 ± 0.007 [0.856, 0.897]	0.707 ± 0.123 [0.594, 0.779]	0.906 ± 0.021 [0.883, 0.923]	0.790 ± 0.008 [0.766, 0.816]	0.660 ± 0.032 [0.633, 0.700]	0.633 ± 0.037 [0.600, 0.673]	**0.919 ± 0.026 [0.894, 0.946]**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.582 ± 0.006 [0.559, 0.612]	0.535 ± 0.023 [0.511, 0.572]	0.426 ± 0.003 [0.422, 0.434]	0.664 ± 0.008 [0.645, 0.690]	0.739 ± 0.009 [0.720, 0.767]	0.637 ± 0.012 [0.618, 0.666]	0.381 ± 0.006 [0.372, 0.396]	**0.717 ± 0.021 [0.692, 0.755]**
	Llama-3-8B	0.725 ± 0.008 [0.703, 0.754]	0.726 ± 0.014 [0.705, 0.754]	0.600 ± 0.016 [0.578, 0.629]	0.725 ± 0.008 [0.705, 0.750]	0.767 ± 0.011 [0.745, 0.795]	0.680 ± 0.026 [0.653, 0.716]	0.544 ± 0.018 [0.526, 0.572]	**0.784 ± 0.003 [0.763, 0.812]**
	Mistral-Small-24B	0.830 ± 0.011 [0.808, 0.850]	0.816 ± 0.010 [0.794, 0.843]	0.588 ± 0.029 [0.561, 0.621]	0.834 ± 0.016 [0.812, 0.856]	0.874 ± 0.005 [0.857, 0.892]	0.687 ± 0.025 [0.662, 0.724]	0.599 ± 0.023 [0.574, 0.632]	**0.867 ± 0.028 [0.841, 0.898]**
AmbigQA	Llama-3.2-3B	0.547 ± 0.014 [0.530, 0.574]	0.544 ± 0.015 [0.527, 0.569]	0.568 ± 0.016 [0.549, 0.593]	0.574 ± 0.031 [0.547, 0.607]	0.531 ± 0.029 [0.505, 0.564]	0.567 ± 0.024 [0.542, 0.597]	**0.598 ± 0.028 [0.572, 0.630]**	0.535 ± 0.013 [0.518, 0.560]
	Llama-3-8B	0.547 ± 0.013 [0.533, 0.572]	0.561 ± 0.011 [0.544, 0.586]	0.517 ± 0.012 [0.507, 0.532]	0.495 ± 0.015 [0.479, 0.520]	0.532 ± 0.035 [0.505, 0.566]	0.573 ± 0.024 [0.549, 0.604]	**0.586 ± 0.026 [0.565, 0.617]**	0.556 ± 0.014 [0.540, 0.582]
	Mistral-Small-24B	0.556 ± 0.026 [0.536, 0.587]	0.564 ± 0.016 [0.545, 0.592]	0.515 ± 0.012 [0.505, 0.530]	0.468 ± 0.007 [0.458, 0.486]	0.559 ± 0.019 [0.540, 0.589]	0.583 ± 0.018 [0.563, 0.610]	**0.594 ± 0.022 [0.573, 0.623]**	0.549 ± 0.017 [0.530, 0.576]

Table 10: AUPR with 95% bootstrap confidence intervals for the additional UQ baselines. Each cell reports the mean ± sample standard deviation over five data-split seeds, with the interval in brackets. OOD and ambiguous prompts form the positive class. EPN is trained with AS-UCE. Blue and light blue cells mark the best and runner-up means in each row, where EPN enters with EU for OOD detection and with AU for ambiguity detection.

Dataset	Backbone	KLE	EigV	CoCoA		Semantic Density	EPN	
				MaxProb	PPL		D (AU)	V (EU)
Alcuna-TriviaQA	Llama-3.2-3B	**0.651 ± 0.013 [0.628, 0.680]**	0.619 ± 0.011 [0.600, 0.650]	0.568 ± 0.011 [0.552, 0.594]	0.539 ± 0.014 [0.523, 0.566]	0.549 ± 0.013 [0.532, 0.576]	0.372 ± 0.025 [0.357, 0.404]	0.640 ± 0.021 [0.613, 0.679]
	Llama-3-8B	**0.826 ± 0.006 [0.806, 0.847]**	0.791 ± 0.016 [0.768, 0.823]	0.775 ± 0.021 [0.752, 0.809]	0.739 ± 0.018 [0.716, 0.772]	0.764 ± 0.018 [0.740, 0.797]	0.540 ± 0.007 [0.526, 0.565]	0.812 ± 0.029 [0.784, 0.851]
	Mistral-Small-24B	0.901 ± 0.010 [0.885, 0.918]	0.891 ± 0.016 [0.871, 0.914]	0.907 ± 0.007 [0.891, 0.924]	0.884 ± 0.006 [0.865, 0.906]	0.883 ± 0.009 [0.865, 0.903]	0.633 ± 0.037 [0.600, 0.673]	**0.919 ± 0.026 [0.894, 0.946]**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.557 ± 0.005 [0.538, 0.581]	0.553 ± 0.007 [0.535, 0.579]	0.552 ± 0.024 [0.528, 0.585]	0.512 ± 0.024 [0.491, 0.543]	0.522 ± 0.025 [0.502, 0.556]	0.381 ± 0.006 [0.372, 0.396]	**0.717 ± 0.021 [0.692, 0.755]**
	Llama-3-8B	0.734 ± 0.008 [0.713, 0.761]	0.746 ± 0.011 [0.722, 0.774]	0.737 ± 0.014 [0.715, 0.765]	0.709 ± 0.017 [0.685, 0.738]	0.741 ± 0.016 [0.717, 0.770]	0.544 ± 0.018 [0.526, 0.572]	**0.784 ± 0.003 [0.763, 0.812]**
	Mistral-Small-24B	0.825 ± 0.009 [0.804, 0.846]	0.854 ± 0.009 [0.834, 0.873]	0.839 ± 0.005 [0.818, 0.859]	0.806 ± 0.003 [0.784, 0.830]	0.827 ± 0.010 [0.806, 0.851]	0.599 ± 0.023 [0.574, 0.632]	**0.867 ± 0.028 [0.841, 0.898]**
AmbigQA	Llama-3.2-3B	0.532 ± 0.008 [0.516, 0.553]	0.534 ± 0.013 [0.517, 0.559]	0.533 ± 0.010 [0.517, 0.556]	0.535 ± 0.017 [0.519, 0.562]	0.529 ± 0.016 [0.513, 0.555]	**0.598 ± 0.028 [0.572, 0.630]**	0.535 ± 0.013 [0.518, 0.560]
	Llama-3-8B	0.545 ± 0.012 [0.530, 0.566]	0.550 ± 0.010 [0.534, 0.574]	0.553 ± 0.012 [0.536, 0.578]	0.556 ± 0.020 [0.538, 0.584]	0.554 ± 0.015 [0.537, 0.580]	**0.586 ± 0.026 [0.565, 0.617]**	0.556 ± 0.014 [0.540, 0.582]
	Mistral-Small-24B	0.549 ± 0.019 [0.531, 0.575]	0.551 ± 0.021 [0.533, 0.581]	0.564 ± 0.021 [0.546, 0.594]	0.565 ± 0.020 [0.545, 0.594]	0.561 ± 0.024 [0.543, 0.592]	**0.594 ± 0.022 [0.573, 0.623]**	0.549 ± 0.017 [0.530, 0.576]

G.4 Additional Results on Decomposition Validity

Cross-evaluation of U_{ale} and U_{epi}. We cross-evaluate U_{ale} and U_{epi} on AmbigQA to assess their task specialization. The modest dissonance AUROC on the original labels may partly reflect a mismatch between human-annotated and backbone-specific ambiguity, where a question with multiple human interpretations may elicit semantically identical responses from a particular backbone (see Appendix F). To examine this mismatch, we exclude human-labeled ambiguous questions whose sampled responses form a single semantic cluster under EPN and evaluate EPN, ICE, and EoE on the same retained questions. Table 11 reports the mean AUROCs over five seeds. We compute the Euclidean distance from this mean pair to the ambiguity ideal $(1.0, 0.5)$. EPN achieves the highest aleatoric AUROC and the smallest distance for all three backbones, showing the best ambiguity specialization among the evaluated decomposition methods on this diagnostic.

Table 11: Cross-evaluation on retained AmbigQA questions. AUROC entries are means over five seeds. Blue cells indicate the highest aleatoric AUROC and smallest distance within each backbone.

Backbone	Method	U_{ale} AUROC ↑	U_{epi} AUROC	Distance to ideal ↓
Llama-3.2-3B	EPN w/ AS-UCE	**0.600**	0.512	**0.400**
	ICE	0.537	0.487	0.463
	EoE	0.469	0.511	0.531
Llama-3-8B	EPN w/ AS-UCE	**0.683**	0.618	**0.338**
	ICE	0.594	0.571	0.412
	EoE	0.550	0.596	0.460
Mistral-Small-24B	EPN w/ AS-UCE	**0.745**	0.625	**0.284**
	ICE	0.648	0.592	0.364
	EoE	0.611	0.607	0.403

Uncertainty component dependence. We compute the Spearman correlation between U_{ale} and U_{epi} separately for each seed and report the mean and sample standard deviation of its absolute value $|\rho_S|$ across five seeds. AmbigQA uses the common retained questions in Table 11; OOD results use each run's evaluated test questions. Table 12 shows that EPN achieves the lowest mean absolute

correlation in eight of nine settings, covering all three retained AmbigQA settings and five of six OOD settings. With equal weight across the nine settings, EPN reduces the mean absolute Spearman correlation by 45.7% and 54.2% relative to ICE and EoE, respectively. These results show weaker rank association between EPN's two uncertainty components.

Table 12: Absolute Spearman correlation $|\rho_S|$ between U_{ale} and U_{epi} (lower indicates weaker rank association). Entries report mean $\pm$ sample standard deviation across five random seeds. AmbigQA uses the retained questions.

Dataset	Backbone	EPN w/ AS-UCE	ICE	EoE
AmbigQA	Llama-3.2-3B	**0.157±0.128**	0.612±0.043	0.504±0.102
	Llama-3-8B	**0.355±0.144**	0.656±0.082	0.616±0.061
	Mistral-Small-24B	**0.351±0.197**	0.531±0.071	0.606±0.070
ALCUNA–TriviaQA	Llama-3.2-3B	**0.059±0.023**	0.610±0.020	0.677±0.029
	Llama-3-8B	0.306±0.033	**0.138±0.049**	0.797±0.018
	Mistral-Small-24B	**0.464±0.014**	0.657±0.270	0.685±0.034
TemporalOOD–TriviaQA	Llama-3.2-3B	**0.076±0.019**	0.442±0.020	0.642±0.034
	Llama-3-8B	**0.351±0.055**	0.760±0.021	0.754±0.033
	Mistral-Small-24B	**0.619±0.020**	0.634±0.136	0.697±0.038

Controlled evaluation of uncertainty components. We evaluate EPN trained by AS-UCE using a paired 2×2 intervention on 150 AmbigQA questions, crossing the original ambiguous wording or a gold disambiguated rewrite with supplied or withheld supporting context, yielding 600 prompts per backbone. Supported and unsupported prompts represent having and lacking the information needed to answer the question, respectively. We replace the original answers with made-up answer labels, where supported prompts provide the label for each interpretation, while unsupported prompts withhold these labels and state that the information is unavailable. Each question keeps the same labels across all four versions, and its clear and ambiguous versions receive the same information within each support condition. We measure the effect of withholding support on vacuity as $\Delta_K U_{\text{epi}} = U_{\text{epi}}^{\text{unsupported}} - U_{\text{epi}}^{\text{supported}}$ and the effect of ambiguous wording on dissonance as $\Delta_A U_{\text{ale}} = U_{\text{ale}}^{\text{ambiguous}} - U_{\text{ale}}^{\text{clear}}$. Table 13 reports mean paired differences and 95% confidence intervals from 10,000 bootstrap replicates that resample base questions while preserving their four variants. For Llama-3-8B and Mistral-Small-24B, withholding support increases vacuity under both wording conditions, and ambiguous wording increases dissonance when support is supplied. This ambiguity effect on dissonance disappears once support is withheld. Llama-3.2-3B shows smaller effects throughout. These results demonstrate sensitivity to supplied context and ambiguity, with dissonance's ambiguity response depending on the availability of supporting information. This pattern aligns with our earlier observation of uncertainty overriding, where epistemic uncertainty can mask the ambiguity signal captured by dissonance.

Table 13: Controlled evaluation of EPN trained by AS-UCE on selected paired questions per backbone. Entries report the mean paired difference, with its 95% paired-bootstrap confidence interval below. Positive $\Delta_K U_{\text{epi}}$ indicates higher vacuity when support is withheld; positive $\Delta_A U_{\text{ale}}$ indicates higher dissonance under ambiguous wording.

Backbone	$\Delta_K U_{\text{epi}}$		$\Delta_A U_{\text{ale}}$	
	Ambiguous	Clear	Supported	Unsupported
Llama-3.2-3B	$+0.039$ $[0.005, 0.072]$	$+0.028$ $[-0.009, 0.064]$	$+0.026$ $[-0.004, 0.056]$	$+0.016$ $[-0.011, 0.042]$
Llama-3-8B	$+0.143$ $[0.104, 0.180]$	$+0.142$ $[0.101, 0.180]$	$+0.043$ $[0.008, 0.079]$	-0.028 $[-0.058, 0.000]$
Mistral-Small-24B	$+0.244$ $[0.205, 0.282]$	$+0.276$ $[0.233, 0.319]$	$+0.048$ $[0.014, 0.082]$	-0.000 $[-0.024, 0.024]$

G.5 Pseudo-OOD Regularization Ablation

This ablation compares EPN trained with AS-UCE against EPN trained with Reg-AS-UCE, which adds the pseudo-OOD regularizer of Appendix C with $\lambda = 0.25$. The pseudo-OOD prompts come from an auxiliary source that differs from the test OOD source, so the test OOD source stays held out. The comparison covers the two OOD datasets. Table 14 reports the results for all four backbones, where the AS-UCE columns repeat the EPN results of Table 1. Reg-AS-UCE improves both AUROC and AUPR in all eight settings, by 1.2 and 2.6 percentage points on average. The gain is larger on ALCUNA, with an average AUROC increase of 1.7 points against 0.7 on TemporalOOD.

Table 14: Pseudo-OOD regularization ablation results from a single run. EPN is trained with AS-UCE or with Reg-AS-UCE ($\lambda = 0.25$), and vacuity is the OOD score in both cases. Bold marks the better value in each pair.

Dataset	Backbone	EPN w/ AS-UCE		EPN w/ Reg-AS-UCE	
		AUROC	AUPR	AUROC	AUPR
Alcuna-TriviaQA	Llama-3.2-3B	0.817	0.694	**0.835**	**0.739**
	Llama-3-8B	0.897	0.855	**0.905**	**0.871**
	Mistral-Small-24B	0.944	0.930	**0.963**	**0.960**
	Qwen3-30B-A3B	0.879	0.806	**0.902**	**0.868**
TemporalOOD-TriviaQA	Llama-3.2-3B	0.876	0.776	**0.889**	**0.804**
	Llama-3-8B	0.829	0.748	**0.831**	**0.749**
	Mistral-Small-24B	0.895	0.872	**0.899**	**0.877**
	Qwen3-30B-A3B	0.854	0.806	**0.862**	**0.828**

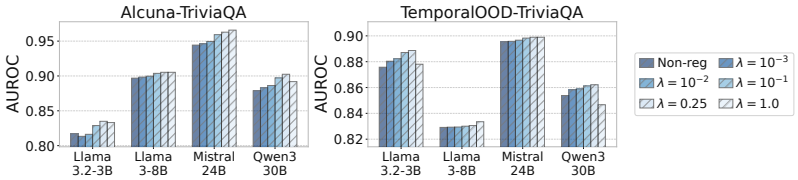


Figure 7: Sensitivity of OOD-detection AUROC to the optional pseudo-OOD regularizer.

Similar to prior work [29, 41, 46], Figure 7 shows that pseudo-OOD regularization improves OOD detection when its strength is properly chosen. The decline in mean AUROC from $\lambda = 0.25$ to 1 indicates that over-weighting the OOD objective can degrade the learned vacuity ranking, confirming its role as an auxiliary regularizer.

G.6 Loss Ablation

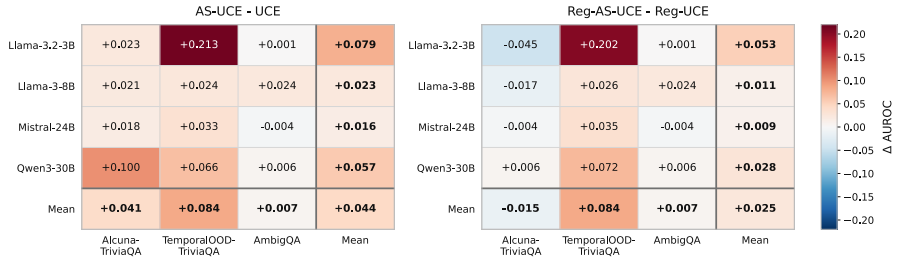


Figure 8: Heatmap comparing AS-UCE and UCE with and without pseudo-OOD regularization. Each cell shows the task-relevant AUROC difference $\Delta\text{AUROC}_{\text{loss}} = \text{AUROC}_{\text{AS-UCE}} - \text{AUROC}_{\text{UCE}}$, where positive values favor AS-UCE. In the Reg-AS-UCE versus Reg-UCE panel, only the eight OOD cells use pseudo-OOD regularization, the four AmbigQA cells repeat the unregularized reference comparison.

Comparison between AS-UCE and UCE. Figure 8 shows that AS-UCE achieves a higher average task-relevant AUROC than UCE in both panels. Without regularization, AS-UCE improves over UCE in 11 of 12 backbone and dataset combinations, with an average gain of 0.044. In the second panel, pseudo-OOD regularization is applied on the two OOD datasets, and AS-UCE leads in 8 of 12 combinations with an average gain of 0.025. The largest gains occur in OOD detection, especially on TemporalOOD, where AS-UCE and Reg-AS-UCE both improve by 0.084 on average. Ambiguity detection shows a smaller average gain of 0.007. The gap narrows in the regularized setting, which suggests that pseudo-OOD supervision partially compensates for the missing unsupported class in UCE.

G.7 Aleatoric Uncertainty Measures

Table 15: Ablation of aleatoric uncertainty measures for AmbigQA ambiguity detection. Blue cells indicate the best AUROC and best AUPR among the candidate measures.

	U_{ale}		$H(\hat{\mathbf{p}})$		$1 - \max_k \hat{\mathbf{p}}_k$	
Backbone	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Llama-3.2-3B	**0.528**	**0.565**	0.451	0.519	0.458	0.527
Llama-3-8B	**0.533**	**0.570**	0.473	0.514	0.472	0.515
Mistral-Small-24B	**0.533**	**0.609**	0.469	0.508	0.473	0.511
Qwen3-30B-A3B	**0.553**	**0.606**	0.497	0.521	0.497	0.519

Aleatoric uncertainty measures. We compare aleatoric uncertainty measures used in prior work [20, 46] using the same retained semantic evidence. Entropy and least confidence use the projected probability vector $\hat{\mathbf{p}}(x) = \mathbf{1}\frac{u(x)}{K} + \mathbf{b}(x)$, whereas dissonance is computed from the retained semantic belief mass. Specifically, we evaluate dissonance U_{ale}, entropy $H(\hat{\mathbf{p}})$, and least confidence $1 - \max_k \hat{\mathbf{p}}_k$. Table 15 shows higher AUROC and AUPR for dissonance on all evaluated backbones. Entropy measures the dispersion of the projected probabilities, and least confidence depends only on the largest one. Dissonance instead targets the conflict among retained semantic beliefs, which makes it the suitable choice for ambiguity detection.

G.8 Supervision Signal Robustness

Judge sensitivity. We examine the sensitivity of EPN to correctness supervision using four instruction-tuned judges with different model capacities, including GPT-5-mini, Mistral-24B, Llama-8B, and Llama-3B. Each judge evaluates generated responses against reference answers to provide the correctness supervision used to construct the targets in $\mathbf{y}^+$ for training and validation. Table 16 reports task-relevant AUROC and AUPR across five random seeds, using vacuity for OOD detection and dissonance for ambiguity detection. We also consider same-backbone supervision, where the backbone itself serves as the correctness judge. Replacing GPT-5-mini with Mistral-24B or Llama-8B changes mean AUROC by at most 1.8 percentage points across the nine settings, suggesting limited sensitivity to these two judge substitutions. The smaller Llama-3B judge also provides competitive supervision, improving ALCUNA–TriviaQA AUROC by 4.0 and 4.4 percentage points for the Llama-3.2-3B and Llama-3-8B backbones, respectively. However, its performance decreases on TemporalOOD–TriviaQA with Llama-3.2-3B by 6.8 percentage points. Overall, EPN is stable under the two mid-sized judges, while the smallest judge provides competitive supervision with larger variation across datasets.

Table 16: Judge sensitivity of EPN trained by AS-UCE. Each cell lists task-relevant AUROC (top) and AUPR (bottom), with mean ± sample standard deviation across five random seeds. † marks same-backbone judging.

Dataset	Backbone	GPT-5-mini judge	Mistral-24B judge	Llama-8B judge	Llama-3B judge
ALCUNA–TriviaQA	Llama-3.2-3B	0.813±0.007 0.640±0.021	0.824±0.008 0.649±0.021	0.816±0.009 0.643±0.015	0.853±0.012 0.700±0.028†
	Llama-3-8B	0.880±0.016 0.812±0.029	0.882±0.006 0.814±0.016	0.878±0.012 0.810±0.022†	0.924±0.005 0.880±0.008
	Mistral-Small-24B	0.936±0.016 0.919±0.026	0.926±0.001 0.904±0.007†	0.926±0.003 0.903±0.009	0.927±0.005 0.885±0.011
TemporalOOD–TriviaQA	Llama-3.2-3B	0.858±0.006 0.717±0.021	0.841±0.012 0.688±0.015	0.840±0.012 0.687±0.017	0.790±0.010 0.619±0.013†
	Llama-3-8B	0.855±0.007 0.784±0.003	0.852±0.008 0.781±0.008	0.848±0.006 0.773±0.005†	0.850±0.011 0.780±0.016
	Mistral-Small-24B	0.897±0.015 0.867±0.028	0.891±0.011 0.860±0.017†	0.914±0.003 0.903±0.006	0.894±0.009 0.844±0.012
AmbigQA	Llama-3.2-3B	0.535±0.025 0.599±0.028	0.535±0.025 0.599±0.028	0.541±0.021 0.603±0.024	0.542±0.021 0.605±0.024†
	Llama-3-8B	0.530±0.009 0.586±0.026	0.530±0.009 0.587±0.026	0.541±0.019 0.595±0.031†	0.538±0.017 0.592±0.027
	Mistral-Small-24B	0.541±0.017 0.597±0.021	0.541±0.017 0.597±0.022†	0.539±0.016 0.594±0.023	0.539±0.016 0.594±0.021

Controlled label noise. We further assess robustness to supervision errors by randomly flipping 10%, 20%, or 40% of the binary correctness labels used to construct $\mathbf{y}^+$ and retraining EPN from scratch. Table 17 reports the results on both OOD benchmarks across four random seeds, using the corresponding uncorrupted runs as the reference. Averaged equally across the six backbone–dataset combinations, the absolute AUROC reductions are 1.09 and 2.35 percentage points at 10% and 20% noise, respectively, indicating limited average degradation under moderate supervision noise in this experiment. At 40% noise, the average reduction increases to 4.71 percentage points. In general, EPN preserves most of its OOD detection performance under moderate correctness-label corruption, while its performance degrades as the noise level increases.

Table 17: Controlled correctness-label noise for EPN trained by AS-UCE on OOD detection. Each cell lists vacuity AUROC (top) and AUPR (bottom), each as the mean ± sample standard deviation across four random seeds. The 0% column uses the corresponding uncorrupted supervision.

Dataset	Backbone	0% noise	10% noise	20% noise	40% noise
ALCUNA–TriviaQA	Llama-3.2-3B	0.814±0.008 0.632±0.015	0.809±0.005 0.624±0.006	0.793±0.014 0.602±0.026	0.787±0.021 0.608±0.027
	Llama-3-8B	0.874±0.004 0.800±0.007	0.874±0.012 0.800±0.023	0.870±0.010 0.796±0.020	0.868±0.008 0.807±0.025
	Mistral-Small-24B	0.929±0.006 0.909±0.015	0.925±0.007 0.896±0.010	0.922±0.005 0.889±0.021	0.918±0.003 0.882±0.026
TemporalOOD–TriviaQA	Llama-3.2-3B	0.856±0.005 0.709±0.010	0.827±0.020 0.671±0.027	0.790±0.042 0.632±0.042	0.704±0.030 0.554±0.030
	Llama-3-8B	0.858±0.004 0.785±0.003	0.845±0.004 0.764±0.012	0.836±0.003 0.749±0.007	0.816±0.013 0.737±0.009
	Mistral-Small-24B	0.915±0.004 0.907±0.008	0.899±0.008 0.874±0.009	0.893±0.012 0.868±0.020	0.870±0.016 0.844±0.028

G.9 Evidence Retention Granularity

While prompt-level EPN uses a shared evidence-retention factor, a response-level variant can condition evidence retention on each generated response. The prompt-level probe uses the hidden state at the token immediately before generation, whereas the response-level retention probe uses the hidden state at the second-to-last token of each sampled response [19]. The latter can therefore condition retention on the realized response rather than on a prompt representation shared across responses. We examine whether predicting evidence retention separately for each sampled response improves uncertainty estimation. Prompt-level EPN predicts a shared factor $q(x)$ from the prompt representation, while response-level EPN predicts $q(x, r^{(m)})$ from the representation of each sampled response $r^{(m)}$. Both variants use AS-UCE and correctness supervision from Mistral-Small-24B. Table 18 reports results across three backbones and three datasets over five random seeds. On AmbigQA, the two variants are nearly tied for all three backbones. For OOD detection, response-level EPN has similar mean AUROC to prompt-level EPN on TemporalOOD with the two Llama backbones and scores lower in the other four settings, by absolute vacuity AUROC gaps of 1.8 to 6.2 percentage points. The response-level variant is more expressive, but it must estimate a separate retention factor from every sampled response, and under matched supervision this flexibility does not yield a consistent gain in the evaluated settings. We therefore keep the shared prompt-level design, which reads the prompt representation once, at the token before generation, and predicts a single factor rather than one for each sampled response. This comparison covers short-form QA only. In long-form generation or multi-step reasoning, the realized response may carry uncertainty signals that the prompt representation alone cannot provide, so response-conditioned retention could become more useful there.

Table 18: Prompt-level versus response-level EPN trained by AS-UCE. Entries report mean ± sample standard deviation across five random seeds. OOD detection uses vacuity U_{epi}, and AmbigQA uses dissonance U_{ale}. The paired difference $\Delta = \text{AUROC}_{\text{response}} - \text{AUROC}_{\text{prompt}}$ favors prompt-level EPN when negative.

Dataset	Backbone	Prompt-level EPN	Response-level EPN	Paired Δ
ALCUNA–TriviaQA	Llama-3.2-3B	0.824±0.008	0.762±0.006	−0.062±0.011
	Llama-3-8B	0.882±0.006	0.854±0.006	−0.028±0.009
	Mistral-Small-24B	0.926±0.001	0.909±0.007	−0.018±0.008
TemporalOOD–TriviaQA	Llama-3.2-3B	0.841±0.012	0.845±0.009	+0.004±0.011
	Llama-3-8B	0.852±0.008	0.850±0.014	−0.002±0.010
	Mistral-Small-24B	0.891±0.011	0.870±0.002	−0.021±0.011
AmbigQA	Llama-3.2-3B	0.535±0.025	0.536±0.025	+0.000±0.001
	Llama-3-8B	0.530±0.009	0.532±0.009	+0.002±0.003
	Mistral-Small-24B	0.541±0.017	0.540±0.016	−0.001±0.003

G.10 Correctness Prediction within OOD Datasets

In this section, we examine whether EPN provides information about answer reliability beyond distinguishing ID from OOD sources. EPN uses AS-UCE trained on ID data, with the test OOD source fully held out. ALCUNA introduces synthetic entities, while TemporalOOD contains events after the backbone's knowledge cutoff. Unfamiliarity with these questions is part of the intended OOD signal with lack of knowledge and does not necessarily constitute a spurious shortcut. Table 19 reports answer accuracy separately for TriviaQA and each OOD source across five random data-split seeds. Accuracy decreases from 0.539–0.831 on TriviaQA to 0.050–0.155 on ALCUNA, and from 0.540–0.830 on TriviaQA to 0.037–0.178 on TemporalOOD, indicating that both OOD sources pose substantial answering challenges for the backbones.

Table 19: Answer accuracy by data source. Entries report mean ± sample standard deviation across five random seeds. TriviaQA and OOD accuracy are evaluated separately within each benchmark.

Benchmark	Backbone	TriviaQA accuracy	OOD accuracy
ALCUNA–TriviaQA	Llama-3.2-3B	0.539±0.041	0.155±0.036
	Llama-3-8B	0.703±0.012	0.070±0.042
	Mistral-Small-24B	0.831±0.006	0.050±0.031
TemporalOOD–TriviaQA	Llama-3.2-3B	0.540±0.041	0.037±0.009
	Llama-3-8B	0.697±0.016	0.138±0.008
	Mistral-Small-24B	0.830±0.007	0.178±0.017

Next, we evaluate correctness prediction exclusively within each OOD source, treating incorrect answers as the positive class. Here, dataset or domain identity is therefore fixed and cannot determine the ranking. Table 20 compares uncertainty scores from EPN, SE, SelfCheckGPT, ICE, and EoE. EPN's vacuity U_{epi} achieves AUROCs of 0.899, 0.772, and 0.781 on TemporalOOD for Llama-3.2-3B, Llama-3-8B, and Mistral-Small-24B, respectively, showing that it captures answer reliability within a fixed data source. For ALCUNA, the corresponding AUROCs are lower than chance, which is due to near-saturated vacuity and the small proportion of correct answers. Together, the results demonstrate that strong source-level OOD detection does not necessarily translate into reliable within-source correctness prediction, although vacuity remains informative for detecting incorrect answers on TemporalOOD. This distinguishes the objectives of correctness prediction and OOD detection.

Table 20: Correctness-prediction AUROC within each OOD source, with incorrect answers as the positive class. Each cell reports the mean (top) and ± sample standard deviation (bottom) across five random data-split seeds. EPN uses AS-UCE and vacuity U_{epi}; both components are reported for ICE and EoE.

OOD source	Backbone	SE	SelfCheck GPT	ICE U_{ale}	ICE U_{epi}	EoE U_{ale}	EoE U_{epi}	EPN w/ AS-UCE U_{epi}
ALCUNA	Llama-3.2-3B	0.657 ±0.031	0.686 ±0.050	0.630 ±0.026	0.534 ±0.024	0.547 ±0.067	0.540 ±0.044	0.356 ±0.032
	Llama-3-8B	0.444 ±0.089	0.462 ±0.130	0.527 ±0.055	0.362 ±0.025	0.626 ±0.061	0.622 ±0.054	0.411 ±0.046
	Mistral-Small-24B	0.430 ±0.131	0.414 ±0.072	0.532 ±0.038	0.496 ±0.036	0.650 ±0.024	0.564 ±0.115	0.490 ±0.074
TemporalOOD	Llama-3.2-3B	0.575 ±0.017	0.493 ±0.042	0.418 ±0.074	0.621 ±0.050	0.765 ±0.108	0.677 ±0.145	0.899 ±0.027
	Llama-3-8B	0.740 ±0.022	0.750 ±0.022	0.765 ±0.038	0.811 ±0.037	0.812 ±0.009	0.757 ±0.019	0.772 ±0.016
	Mistral-Small-24B	0.785 ±0.029	0.804 ±0.027	0.632 ±0.033	0.748 ±0.029	0.812 ±0.024	0.684 ±0.069	0.781 ±0.030

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the Abstract and Section 1. These claims are developed in Section 4, analyzed in Section 5, and supported empirically in Section 6, with scope and assumptions discussed in Section 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses the main scope and assumptions of the work, including EPN's backbone requirements, the evaluation scope, and directions for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 5 states the assumptions for each theoretical result and gives proof sketches for the UCE distortion and AS-UCE evidence-rejection claims. Full proofs and additional derivations are provided in Appendices B.2 and B.3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The description of the uncertainty decomposition pipeline is well-defined in the main paper. Details to reproduce the experiments and ablation studies can be found in Appendices D.1 and D.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  4.1. If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  4.2. If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  4.3. If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  4.4. We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A code link is provided in the main paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details can be found in Appendices D.1 and D.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report means and sample standard deviations across five data-split seeds for the multi-seed main comparisons. Appendix G.3 reports hierarchical-bootstrap confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute hardware is described in Appendix D.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: N/A

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper motivates the positive impact of more diagnostic uncertainty estimates for improving the reliability of free-form LLM generation, but the current draft does not include an explicit discussion of negative societal impacts or misuse risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The citation list includes all original papers whose code is used in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We included a document along with the code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not contain crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not contain crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Sections 4 and 6 describe the backbone LLMs and their role in the uncertainty-decomposition pipeline. Appendices D.1, D.3, and G.8 describe LLM use for dataset construction, correctness supervision, and baseline clarification.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.