Credal Concept Bottleneck Models for Epistemic–Aleatoric Uncertainty Decomposition

Tanmoy Mukherjee[1], Thomas Bailleux,[1], Pierre Marquis[1,2], Zied Bouraoui[1]
[1] Univ. Artois, CNRS, CRIL, France, [2] Institut Universitaire de France
{mukherjee,bailleux,marquis,bouraoui}@cril.fr

Abstract

Concept Bottleneck Models (CBMs) predict through human-interpretable concepts, but they typically output point concept probabilities that conflate epistemic uncertainty (reducible model underspecification) with aleatoric uncertainty (irreducible input ambiguity). This makes concept-level uncertainty hard to interpret and, more importantly, hard to act upon. We introduce CREDENCE (Credal Ensemble Concept Estimation), a CBM framework that decomposes concept uncertainty by construction. CREDENCE represents each concept as a credal prediction (a probability interval), derives epistemic uncertainty from disagreement across diverse concept heads, and estimates aleatoric uncertainty via a dedicated ambiguity output trained to match annotator disagreement when available. The resulting signals support prescriptive decisions: automate low-uncertainty cases, prioritize data collection for high-epistemic cases, route high-aleatoric cases to human review, and abstain when both are high. Across several tasks, we show that epistemic uncertainty is positively associated with prediction errors, whereas aleatoric uncertainty closely tracks annotator disagreement, providing guidance beyond error correlation. Our implementation is available at the following link: https://github.com/Tankiit/Credal_Sets/tree/ensemble-credal-cbm

1 Introduction

In high-stakes settings, such as healthcare, legal analysis, and financial advising, NLP systems are increasingly used to support consequential decisions. In these contexts, it is not enough to output a prediction; practitioners must also assess when the system is reliable and, crucially, *why* it may be unreliable. A key distinction lies between *epistemic uncertainty* (what the model has not yet learned but could learn with additional evidence) and *aleatoric uncertainty* (irreducible ambiguity or noise in the input) (Kendall and Gal, 2017; Shaker and Hüllermeier, 2020). Existing models face challenges when dealing with uncertainty (Hüllermeier et al., 2022; Shaker and Hüllermeier, 2020), as they generate a single confidence score that merges various types of uncertainty, leading to a uniform approach for managing all uncertain predictions. Consider a restaurant review classifier that processes the sentence: *"The waiter was efficient, I suppose."* A model might assign a 60% probability that the sentiment about service is positive. However, this single number conflates two qualitatively different situations. The model might be uncertain because it has seen few hedged expressions such as "I suppose" during training (epistemic uncertainty). Alternatively, the text may genuinely express ambivalence, so that competent annotators could disagree even with unlimited training data (aleatoric uncertainty). The distinction is actionable: high epistemic uncertainty suggests targeted data collection or modeling changes, while high aleatoric uncertainty suggests human review, user-facing ambiguity communication, or abstention.

Making uncertainty actionable often requires exposing *what* the model is uncertain about, not only *how much* uncertainty it has. For many NLP decisions, practitioners reason in terms of interpretable attributes—e.g., whether a review expresses praise for *service*, whether a sentence contains *hedging*, or whether a post includes *insults*. This suggests targeting uncertainty at the level of such intermediate attributes. Concept Bottleneck Models (CBMs) (Koh et al., 2020) provide a natural framework for this setting because they explicitly use a set of human-interpretable concepts before producing the final label. However, standard CBMs typically represent each concept as a point probability, which collapses epistemic and aleatoric effects at the concept layer. As a result, concept-level uncertainty becomes difficult to interpret and, more importantly, difficult to act upon: it may indicate which predic-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42961–42980

tions are risky, but not what response is appropriate.

Different uncertainty types should trigger different interventions: high epistemic uncertainty calls for collecting additional evidence or improving the model; high aleatoric uncertainty calls for human oversight or qualified outputs; and high–high cases motivate abstention or escalation. This issue is particularly pronounced in NLP: tasks involving sentiment, toxicity, and emotion are fundamentally subjective, i.e. "ground truth" reflects a composite of disputed human judgments. As a result, distinguishing between epistemic and aleatoric uncertainty is not just a theoretical refinement but a practical requirement for any triage system deployed in production. Without decomposition, uncertain cases are treated uniformly, conflating fixable model limitations with irreducible ambiguity. To address this, we introduce CREDENCE (Credal Ensemble Concept Estimation), a CBM framework that decomposes concept-level uncertainty by construction. The key insight is *structural separation*: epistemic and aleatoric signals must come from different sources to prevent them from collapsing. Concretely, we compute epistemic uncertainty from *disagreement across* ensemble heads and aleatoric uncertainty from a dedicated head trained *within* each example on annotator disagreement.

Our contributions are: (i) a credal CBM formulation that represents concept predictions as probability intervals and supports decisions (trust, abstain, escalate to human); (ii) a structurally separated decomposition of concept uncertainty into epistemic (ensemble disagreement) and aleatoric (predicted ambiguity) components derived from different parameters and avoiding collapse; and (iii) empirical evaluation across several NLP tasks, showing that epistemic uncertainty aligns with prediction errors, whereas aleatoric uncertainty tracks annotator disagreement, yielding actionable guidance beyond error correlation.

2 Related Works

Interpretability in NLP spans a spectrum from local explanations tied to a specific input (e.g., token- or span-level rationales) to global explanations that describe model behavior as higher-level semantic concepts. Our research lies at the intersection of concept-based interpretability and uncertainty quantification: we seek to make uncertainty at the concept level not only something that can be measured but also something that can guide decisions.

Concept Bottleneck Models. CBMs (Koh et al., 2020) route predictions through human-interpretable concepts, enabling inspection and intervention. NLP adaptations include Text Bottleneck Models (Ludan et al., 2024), which discover concepts via LLMs, CLARITY (Bailleux et al., 2025) connecting concepts to rationales, and CB-LLMs (Sun et al., 2025), which align neuron activations with concept scores. Probabilistic extensions (Kim et al., 2023; Collins et al., 2023) introduce stochastic embeddings and study intervention under uncertainty. However, existing CBMs represent concepts as point probabilities, conflating epistemic and aleatoric uncertainty.

Rationales and Local Explanations. Rationale extraction (Lei et al., 2016; Bastings et al., 2019), attention-based explanations (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019), and feature attribution (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017) highlight tokens but do not distinguish *why* a model is uncertain - the same diffuse rationale may reflect model underspecification (epistemic) or linguistic ambiguity (aleatoric).

Uncertainty Estimation Dominant approaches include MC Dropout (Gal and Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017), and evidential methods (Sensoy et al., 2018). Uncertainty quantification has been applied to translation, QA, and classification (Xiao and Wang, 2019; He et al., 2020), with recent work on LLM calibration (Kadavath et al., 2022; Kuhn et al., 2023). The epistemic-aleatoric distinction, studied in vision (Kendall and Gal, 2017; Depeweg et al., 2018), has received relatively less attention in NLP. Recent work by Xiao and Wang (2019) decomposes uncertainty but relies exclusively on ensemble disagreement, while Baan et al. (2023); Ulmer (2024); Fisch et al. (2022) provide surveys of uncertainty sources in language generation. A central limitation of these approaches is that they focus on model outputs rather than on intermediate concept representations, and therefore do not provide concept-level guidance on *how to respond* to each distinct type of uncertainty. We address this gap by operating at the *concept level* within an interpretable pipeline, thereby enabling both principled uncertainty decomposition and targeted, concept-specific interventions.

Imprecise Probabilities and Credal Sets. Credal sets represent uncertainty about prob-

abilities rather than committing to a single estimate (Walley, 1991; Levi, 1980). (Shaker and Hüllermeier, 2020) provide a comprehensive treatment of aleatoric and epistemic uncertainty. Credal sets have been applied to classification with reject option (Corani and Zaffalon, 2008; Zaffalon et al., 2012) and set-valued prediction (Mortier et al., 2022). Sale et al. (2023) recently studied set-valued predictions for reliable uncertainty quantification. We leverage this framework to obtain concept-level uncertainty signals that better separate epistemic and aleatoric effects.

Positioning Prior CBM work focuses on concept-mediated interpretability and interventions, but typically represents concepts as point probabilities and does not separate epistemic and aleatoric uncertainty at the concept layer. Uncertainty quantification in NLP studies error-aware confidence and calibration, but rarely yields concept-level prescriptions about whether to collect data, route to humans, or abstain. We connect these lines by introducing a credal, ensemble-based CBM that structurally separates epistemic disagreement from annotator-derived ambiguity, and we evaluate it in terms of prescriptive routing rather than error correlation alone.

3 Methodology

We introduce CREDENCE, a concept bottleneck model that makes concept-level uncertainty explicit and actionable. The method is built around two ideas: (i) represent each concept prediction as a *probability interval* rather than a single number, and (ii) decompose uncertainty into epistemic and aleatoric uncertainty from *structurally different sources* so they do not collapse into each other.

Background. A CBM predicts through an intermediate concept layer (Koh et al., 2020):

$$\mathbf{x} \xrightarrow{f_{\text{enc}}} \mathbf{h} \xrightarrow{g} \mathbf{c} \xrightarrow{f_{\text{cls}}} \hat{y} \tag{1}$$

where x is the input text, $\mathbf{h} \in \mathbb{R}^d$ is the encoded representation from a pre-trained encoder, $\mathbf{c} = [c_1, \ldots, c_K] \in [0,1]^K$ are K concept probabilities (e.g "service quality is positive", "food quality is positive"), and $\hat{y}$ is the prediction. Each concept layer typically outputs point estimates $\hat{p}_k = P(c_k = 1 \mid x)$. In such a model, a single $\hat{p}_k(x)$ conflates epistemic uncertainty (is the model confused) and aleatoric uncertainty (input ambiguity).

From Point Estimates to Credal Sets. Instead of committing to a single $\hat{p}_k(x)$, we predict an *interval*

$$P(c_k = 1 \mid x) \in [\underline{p}_k, \overline{p}_k] \tag{2}$$

which defines a *credal set* in imprecise probability theory (Walley, 1991). A credal set represents a range of probabilities instead of a single value. Rather than stating "the probability is 0.7," we say "it falls between 0.6 and 0.8." The span of this interval reflects our uncertainty, while its midpoint reflects our best estimate.

Problem Statement. Given an input text x and label space $\mathcal{Y}$, our goal is to produce: (i) concept-level credal predictions $\{[\underline{p}_k(x), \overline{p}_k(x)]\}_{k=1}^K$; (ii) a decomposition of concept uncertainty into an epistemic signal (model confusion) and an aleatoric signal (text ambiguity); and (iii) a prediction $\hat{y}$ (or a set of plausible labels) via decision rules. A complete notation reference appears in Appendix A.

3.1 CREDENCE Architecture

CREDENCE follows a four-stage pipeline: (i) encode the input, (ii) produce multiple concept predictions using diverse lightweight heads, (iii) aggregate these predictions into concept intervals, and (iv) propagate intervals through the classifier. An overview is shown in Figure 1:

$$x \xrightarrow{f_{\text{enc}}} \mathbf{h} \begin{cases} \xrightarrow{\text{Ens.}} \{\hat{p}_h^{(k)}\}_{h=1}^H \to U_{\text{epi}}^{(k)} \\ \xrightarrow{\text{Ale.}} \sigma_\theta^{(k)} \to U_{\text{ale}}^{(k)} \end{cases} \xrightarrow{\text{Agg.}} \mathcal{C}_k \xrightarrow{f_{\text{cls}}} \hat{y} \tag{3}$$

Stage 1:Encoder. A frozen pre-trained model produces the representation $\mathbf{h} = f_{\text{enc}}(x)$. We evaluate with encoder-only models and decoder-only LLMs. Refer to Sec 4 for details on the encoders used. The encoder is *not* fine-tuned; only the concept heads are trained.

Stage 2: Ensemble of Concept Heads Instead of a single concept predictor g, we use H diverse heads $\{g_1, \ldots, g_H\}$. Each head independently predicts all K concepts:

$$\hat{p}_h^{(k)} = \sigma\big(g_h(\mathbf{h})_k\big) \quad \text{for } k = 1, \ldots, K \tag{4}$$

where σ is the sigmoid function. The heads are implemented as LoRA adapters (Hu et al., 2021) - trainable matrices added to the frozen encoder, enabling H diverse predictors without retraining

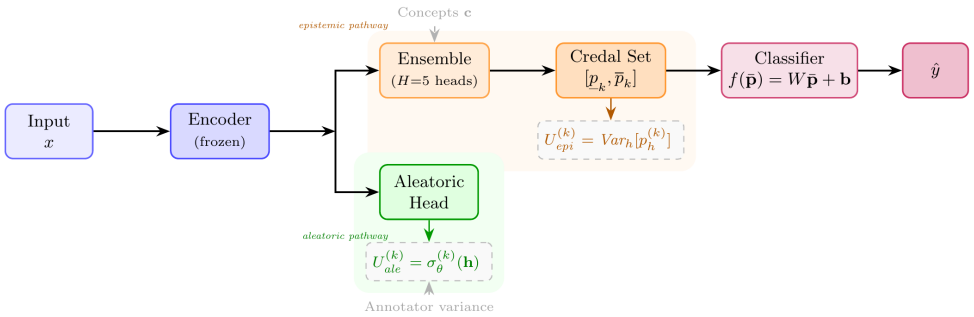

Figure 1: CREDENCE architecture. Epistemic uncertainty emerges from ensemble disagreement (top branch); aleatoric uncertainty is predicted by a dedicated head supervised by annotator variance (bottom branch).

millions of parameters. Each head uses a different LoRA rank $r_h \in \{4, 8, 16, 32, 64\}$: low-rank heads capture broad patterns, while high-rank heads capture finer details. This variation under a fixed backbone ensures that heads see the same input but reach different predictions on the same input, which will be used to quantify *epistemic uncertainty*. Implementation details are in Appendix B.1.

Stage 3: Credal Aggregation. We form credal intervals by taking the minimum and maximum across heads:

$$\mathcal{C}_k = [\underline{p}_k, \bar{p}_k] = \left[\min_h \hat{p}_h^{(k)},\ \max_h \hat{p}_h^{(k)}\right] \quad (5)$$

This min/max construction yields conservative bounds; a narrow interval means heads agree (low epistemic uncertainty); a wide interval means heads disagree (high epistemic uncertainty).

Stage 4: Classification. A linear classifier maps mean concept predictions to labels:

$$\hat{y} = \arg\max_j\ f_{\text{cls}}(\bar{\mathbf{p}})_j \quad \text{where} \quad \bar{p}_k = \frac{1}{H}\sum_{h=1}^{H} \hat{p}_h^{(k)} \quad (6)$$

For ΔAcc evaluation, we use the mean prediction $\bar{\mathbf{p}}$. For uncertainty-aware decisions, we propagate credal bounds through the classifier (see §3.1).

Uncertainty Decomposition We compute two concept-level signals:

$$U_{\text{epi}}^{(k)} = \text{Var}_h\left[p_h^{(k)}\right] \quad \text{(Epistemic)}$$
$$U_{\text{ale}}^{(k)} = \sigma_\theta^{(k)}(\mathbf{h}) \quad \text{(Aleatoric)}$$

where $\sigma_\theta^{(k)}$is a learned prediction of inherent ambiguity trained to predict annotator disagreement when it is available. The key point is that these signals derive from *different parameters*: epistemic uncertainty measures variation *across* heads, while aleatoric uncertainty is predicted *within* each head. Because they have different sources, they cannot collapse into a single score. This architectural guaranty extends the epistemic/aleatoric framework of Shaker and Hüllermeier (2020) to the concept level and instantiates the set-valued prediction approach of Sale et al. (2023) within a CBM pipeline. We validate that epistemic tracks prediction errors and aleatoric tracks annotator disagreement (§4.1).

Credal Classification and Decision Rules Given credal bounds $[\underline{p}_k, \bar{p}_k]$ for each concept, we compute exact logit bounds via interval arithmetic:

$$\underline{\ell}_j = \sum_{k:W_{jk}>0} W_{jk}\underline{p}_k + \sum_{k:W_{jk}<0} W_{jk}\bar{p}_k + b_j \quad (7)$$
$$\bar{\ell}_j = \sum_{k:W_{jk}>0} W_{jk}\bar{p}_k + \sum_{k:W_{jk}<0} W_{jk}\underline{p}_k + b_j \quad (8)$$

These bounds are tight (proof in Appendix C). We use mean predictions $\bar{\mathbf{p}}$ for classification and uncertainty signals for routing decisions (§4). Credal bounds also support principled decision criteria from imprecise probability theory (e.g., **Γ-Maximin**, Maximality); see Appendix E.

Training and Inference. We train the ensemble heads, aleatoric head, and classifier jointly while keeping the encoder frozen. The loss combines three terms: $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_c\mathcal{L}_{\text{concept}} + \lambda_a\mathcal{L}_{\text{ale}}$, where $\mathcal{L}_{\text{task}}$ is cross-entropy on mean predictions, $\mathcal{L}_{\text{concept}}$ supervises concept predictions across all heads, and $\mathcal{L}_{\text{ale}}$ trains the aleatoric head to predict annotator disagreement when available. At test time, we compute predictions from all H heads, aggregate into credal bounds, and derive uncertainty signals; no annotator labels are used. Full training and inference procedures are described in Algorithms 1–2 (Appendix D).

Table 1: Dataset statistics. CEBaB provides explicit concept annotations and "unknown" labels essential for aleatoric validation.

Dataset	Task	$\|C\|$	$\|K\|$	Train	Ann.	Src
CEBaB	Sentiment	3	4	9.8K	✓	Human
SST-2	Sentiment	2	20	67K	–	LLM
GoEmotions	Emotion	28	28	43K	✓	Human
HateXplain	Toxicity	3	2	15K	✓	Human

4 Experiments

We evaluate CREDENCE through three research questions. **RQ1 (Decomposition Validity)**: Does epistemic uncertainty $U_{\text{epi}}^{(k)} = \text{Var}_h[p_h^{(k)}]$ correlate with prediction errors? **RQ2 (Ambiguity detection)**: Does aleatoric uncertainty $U_{\text{ale}}^{(k)} = \sigma_\theta^{(k)}$ correlate with annotator disagreement? **RQ3 (Actionability)**: Does knowing uncertainty *type* enable more effective concept interventions? RQ1–RQ2 validate that our decomposition captures distinct phenomena (model ignorance vs. data ambiguity); RQ3 shows this distinction enables more effective concept interventions.

Datasets. We evaluate three sentiment/emotion/toxicity datasets that include multi-annotator labels, enabling direct validation of aleatoric uncertainty against ground-truth annotator disagreement (Table 1). We additionally include SST-2, for which we obtain LLM-generated annotations to broaden the analysis and support ablations.

For intervention experiments, we focus our analysis on CEBaB (Abraham et al., 2022) for three reasons (i) *explicit concept annotations* (food, service, ambiance, noise) enabling oracle interventions; (ii) *causal structure* where concepts mediate the review-to-rating relationship (Abraham et al., 2022); and (iii) *explicit unknown labels* (52.2% of annotations), providing ground-truth for aleatoric validation.

Model Backbones and Training We evaluate across two model families. *Encoder models* use a frozen encoder with trainable heads: DistilBERT-base (66M parameters), RoBERTa-base (125M), and DeBERTa-v3-base (184M). *LLM models* use LoRA fine-tuning (r=16, α=32): Phi-3-mini (3.8B), Mistral-7B (7B), and Llama-3.1-8B (8B). All configurations use H=5 ensemble heads with diverse dropout rates (0.05–0.25) and pooling strategies (CLS and mean pooling) to encourage prediction diversity. For training, we use AdamW optimizer with lr=10^{-4}, batch size 16, weight decay 0.01. Encoder models train 40 epochs; LLMs train 10 epochs with LoRA. Loss weights: λ_{concept}=1.0, λ_{ale}=0.5. This tests whether decomposition generalizes across model scales. See Appendix B–D for details.

Baselines. We compare against established uncertainty quantification methods: MC Dropout (Gal and Ghahramani, 2016) (50 stochastic forward passes), Deep Ensembles (Lakshminarayanan et al., 2017) (5 independently trained models), Temperature Scaling (Guo et al., 2017) (post-hoc calibration), and Evidential DL (Sensoy et al., 2018) (Dirichlet-based uncertainty). We also include CBM-specific baselines: CBM (Koh et al., 2020), CBM+MC Dropout, CBM+Ensemble, and P-CBM (Kim et al., 2023). All methods are matched for computational budget (5 ensemble members or 50 MC passes). More detail in App F.

Metrics. We report three primary metrics. (i) **Acc**: task accuracy computed from the ensemble mean prediction $\bar{\mathbf{p}}$, as defined in §3.1. This is *not* an accuracy gain; we use a plain label to avoid confusion with the intervention metric below. (ii) $\rho_{\textbf{epi}}$: Spearman correlation between sample-level epistemic uncertainty $U_{\text{epi}} = \frac{1}{K}\sum_k U_{\text{epi}}^{(k)}$ and a binary error indicator (1 if prediction is wrong, 0 otherwise). Positive values indicate that epistemic uncertainty is higher on incorrect predictions. (iii) $\rho_{\textbf{ale}}$: Spearman correlation between $U_{\text{ale}}^{(k)}$ and the per-concept annotator unknown rate. Positive values indicate aleatoric uncertainty tracks human ambiguity. On SST-2, which lacks multi-annotator disagreement labels, ρ_{ale} reflects correlation with the error indicator instead; we expect and observe negative values, confirming that aleatoric does *not* track errors (see §4.1). (iv) **ΔAcc** (intervention tables only): accuracy gain from replacing selected concept predictions with ground-truth values, i.e. $\text{Acc}(\text{corrected}) - \text{Acc}(\text{original})$.

Table 2: Main results across datasets and model scales. **Top**: General UQ and CBM baselines. **Middle**: CREDENCE with encoder backbones (66M–184M params). **Bottom**: CREDENCE with LLM backbones (3.8B–8B params, LoRA). ρ_{epi}: Spearman correlation between epistemic uncertainty and prediction error. ρ_{ale}: correlation with annotator disagreement (SST-2 lacks disagreement labels; values show error correlation). Best per section in **bold**. All $p < 0.001$.

		CEBaB			GoEmotions			HateXplain			SST-2		
	Method	Acc	ρ_{epi}	ρ_{ale}	Acc	ρ_{epi}	ρ_{ale}	Acc	ρ_{epi}	ρ_{ale}	Acc	ρ_{epi}	ρ_{ale}
General	MC Dropout	70.2	.185	.312	47.1	.042	.089	73.8	.178	.234	89.8	.198	-.089
	Deep Ensemble	71.1	.201	.345	48.2	.056	.112	74.5	.195	.267	90.3	.215	-.076
	Temp. Scaling	70.8	.098	.287	47.5	.029	.076	74.1	.112	.198	90.1	.124	-.102
	Evidential DL	69.5	.142	.287	45.8	.031	.078	72.9	.134	.198	89.2	.156	-.095
CBM	Standard CBM	70.4	.121	.298	46.8	.038	.094	73.2	.145	.221	90.6	.238	-.111
	CBM + MC Drop	70.1	.168	.324	46.5	.044	.102	73.0	.162	.245	90.3	.261	-.098
	CBM + Ensemble	71.0	.189	.356	47.9	.052	.118	74.2	.184	.271	90.9	.274	-.085
	P-CBM	70.6	.156	.341	47.2	.048	.108	73.6	.171	.258	90.5	.252	-.095
Encoder	CRED.-DistilBERT	68.2	.251	.742	45.3	.089	.183	72.1	.234	.412	89.0	.316	-.128
	CRED.-RoBERTa	71.6	.287	.785	48.7	.071	.198	74.8	.267	.445	90.4	.342	-.112
	CRED.-DeBERTa	**73.4**	**.302**	**.785**	**51.2**	**.095**	**.198**	**76.3**	**.289**	**.463**	**91.2**	**.361**	-.095
LLM	CRED.-Phi-3	69.8	.268	.723	46.1	.082	.176	73.5	.245	.398	88.6	.325	-.118
	CRED.-Mistral	**72.3**	**.279**	**.758**	**49.4**	**.103**	.185	**75.2**	**.271**	**.431**	**90.2**	**.348**	-.102
	CRED.-Llama-3	71.9	.274	.749	48.9	.091	**.182**	74.9	.263	.427	89.9	.339	**-.098**

4.1 Main Results

Table 2 summarizes results across datasets and model scales. The core finding is *separation*: CREDENCE is the only method achieving high ρ_{ale} where disagreement labels exist (CEBaB: 0.785 vs. best baseline 0.356, a $2.1\times$ improvement), while consistently improving ρ_{epi} over matched baselines (CRED.-RoBERTa vs. CBM+Ensemble: 0.287 vs. 0.189, same backbone, same 5-head budget). This separation enables the downstream decisions in Tables 3 and 4: which concepts to correct and how to route samples in sentiment and toxicity pipelines. Among encoders, DeBERTa achieves the strongest epistemic correlations; LoRA-adapted LLMs are competitive but not superior; Mistral-7B reaches $\rho_{\text{epi}} = 0.348$ on SST-2, slightly below DeBERTa despite a substantially larger scale, consistent with encoder CBMs being more sample-efficient under full concept supervision. We now examine each research question in detail.

RQ1: Does Epistemic Uncertainty track Prediction Errors? Table 2 shows CREDENCE achieves strong epistemic-error correlations across all datasets. Stratifying predictions by correctness reveals a clear pattern: when the model makes an error, ensemble heads disagree; when the model is correct, heads converge. This means epistemic uncertainty can serve as a reliable signal for identifying predictions that should not be trusted. In contrast, aleatoric uncertainty shows no such pattern—incorrect and correct predictions have nearly identical aleatoric values. High aleatoric does not indicate that the model will be wrong; it indicates something else entirely. On SST-2, ρ_{ale} is negative across all methods because the dataset lacks annotator disagreement labels; aleatoric has no ground-truth ambiguity to track, so it correctly shows no systematic association with errors. A negative value does *not* imply an inverse relationship; it reflects the null expectation when no ground-truth ambiguity exists. This validates structural separation: epistemic (ensemble variance) and aleatoric (learned ambiguity) track distinct phenomena. If aleatoric does not track errors, what *does* it capture? We turn to CEBaB, which provides ground-truth annotator disagreement.

RQ2: Does aleatoric uncertainty capture data ambiguity? CEBaB provides explicit "unknown" annotations when annotators could not determine a concept's value (52.2% of labels). On this dataset, aleatoric strongly correlates with annotator disagreement: $\rho_{\text{ale}} = 0.742$–$0.785$ for CREDENCE models (Table 2), far exceeding baselines (best: CBM+Ensemble at $\rho_{\text{ale}} = 0.356$). The correlation strengthens with concept ambiguity: Food (25% unknown, $\rho = 0.72$) $\rightarrow$ Service (45%, $\rho = 0.78$) $\rightarrow$ Ambiance (63%, $\rho = 0.81$) $\rightarrow$ Noise (75%, $\rho = 0.83$). Per-concept breakdown appears in Appendix F.4 Table 11. This means that aleatoric uncertainty captures genuine input ambiguity—cases

Table 3: Concept intervention results. ΔAcc = Acc(corrected)−Acc(original) after replacing top-5 concepts with ground-truth labels.

Dataset	Epistemic	Aleatoric	Ratio
SST-2	+2.1%	+19.1%	9.1×
CEBaB	+6.4%	+18.7%	2.9×
Mean	+4.3%	+18.9%	**4.4**×

where even humans cannot agree. High aleatoric flags text that is inherently unclear, not text that the model misunderstands.

The SST-2/CEBaB Contrast Validates Structural Separation. The opposing behaviours confirm that epistemic and aleatoric capture genuinely distinct phenomena. On *SST-2* (no disagreement labels): epistemic is positively associated with errors ($\rho_{\text{epi}} > 0$); aleatoric shows a small negative association with errors ($\rho_{\text{ale}} < 0$), indicating it is *not* measuring model failure as intended. On *CEBaB* (has disagreement labels): epistemic is again positively associated with errors ($\rho_{\text{epi}} > 0$); aleatoric is strongly associated with human annotator disagreement ($\rho_{\text{ale}} = 0.78$), not with model errors. Note that a negative ρ_{ale} on SST-2 does *not* imply an inverse relationship; it reflects the absence of a systematic link, which is the null expectation when no ground-truth ambiguity signal exists. Likewise, a positive ρ_{ale} on CEBaB reflects a genuine positive association, not merely a non-zero value. If both signals arose from the same source, they would behave identically across datasets. Instead, epistemic consistently tracks model confusion regardless of annotation availability, while aleatoric tracks human disagreement only when such labels exist. This is the structural separation we designed.

RQ3: Does decomposition enable better decisions? RQ1-RQ2 **RQ1RQ2** established that epistemic and aleatoric capture distinct phenomena. We now turn to whether this distinction is actually useful in practice by analyzing two views: (i) targeted concept corrections and (ii) sample routing.

Application 1: Concept Interventions. We replace predicted concepts with ground-truth labels, comparing three strategies: target concepts with the highest epistemic uncertainty, highest aleatoric uncertainty, or random selection.

Table 3 reveals an asymmetry that validates the decomposition: epistemic and aleatoric identify *functionally different* concept sets, so they respond differently to the same correction procedure. This is not a competition between two strategies: there is no "satisfactory" ratio threshold. The relevant question is whether the two uncertainty types identify distinct concepts; the intervention gap confirms they do, consistently across runs (aleatoric: $\pm$0.2pp variance; epistemic: $\pm$2.4pp, Table 12).

Why Aleatoric-Targeted Corrections Work Better? The answer lies in what each uncertainty type captures. High-aleatoric concepts are ambiguous *because they matter*; annotators disagree precisely on concepts that strongly influence the final label. Correcting these resolves genuine decision-relevant ambiguity. High-epistemic concepts, by contrast, reflect model confusion—but this confusion often occurs on rare patterns or edge cases that are weakly connected to the prediction. Fixing what the model struggles with is not the same as fixing what matters for the task. This has direct practical implications: when the annotation budget is limited, prioritize correcting concepts with high aleatoric uncertainty. These are the concepts where human input provides maximum information gain.

Application 2: Sample Routing. Beyond concept corrections, decomposition enables routing entire samples to appropriate handlers. Table 4 shows examples from four quadrants defined by median uncertainty thresholds. The quadrants reveal qualitatively different failure modes that demand different responses:

DATA (56.6% ΔAccuracy): The model fails on clear inputs. These are learnable errors: the text "Lobster Mac & Cheese is incredible. Service was terrible." has unambiguous aspect sentiments, but the model has not seen enough mixed-sentiment examples to aggregate them correctly. *Response: collect more training data.*

REVIEW (85.7% ΔAccuracy): The model succeeds on ambiguous inputs. "Disappointing." is correctly classified as negative, but reasonable humans might disagree about severity or aspect attribution. *Response: flag for human review, not because the model is wrong, but because stakeholders may legitimately disagree.*

Without decomposition, both cases would be labeled simply "uncertain" and receive identical treatment: missing the opportunity for targeted action. The ΔAcc gap between DATA and REVIEW quadrants is invisible to any aggregate uncertainty measure. Full quadrant analysis appears in Appendix G.

Table 4: Representative examples from each uncertainty quadrant (CEBaB).

Text	Pred	U_{epi}	U_{ale}	Interpretation
TRUST (Low Epi, Low Ale) — ΔAcc: 78.8% — *Automate*				
"The service was fantastic and the food was very good."	Pos ✓	.001	.25	Model confident, input clear
DATA (High Epi, Low Ale) — ΔAcc: 56.6% — *Collect training data*				
"Lobster Mac & Cheese is incredible. Service was terrible."	Neg ✗	.017	.50	Model confused, but input is clear
REVIEW (Low Epi, High Ale) — ΔAcc: 85.7% — *Human review*				
"Disappointing."	Neg ✓	.002	.89	Model confident, but humans may disagree
ABSTAIN (High Epi, High Ale) — ΔAcc: 65.3% — *Decline*				
"Been there many times."	Neu ✗	.005	.98	Model confused, input unclear

✓= correct, ✗= error. Extended examples in Appendix H.

Not All Uncertainty the Same Consider two predictions, both 70% uncertain. Standard approaches treat them identically—flag both for review or trust neither. But RQ1–RQ3 show this misses critical information: (i) **Prediction A**: High epistemic, low aleatoric. The model is confused, but the text is clear. A human would easily label this correctly. *Solution: train the model on more similar examples.* (ii) **Prediction B**: Low epistemic, high aleatoric. The model is confident, but the text is genuinely ambiguous; even humans disagree. *Solution: ask a human, as this is a judgment call.* These require opposite responses, yet aggregate uncertainty cannot distinguish them. CREDENCE can, because epistemic (model confusion) and aleatoric (human disagreement) arise from separate components that measure different things. Knowing *why* a prediction is uncertain tells you *what to do about it*.

4.2 Ablation Studies

We ablate key design choices on CEBaB with RoBERTa-base. Table 5 summarizes results, with full ablations in Appendix I, including LoRA rank configurations (Appendix I.3), dropout spacing (Appendix I.4), loss weight sensitivity (Appendix I.6), cross-dataset consistency (Appendix I.7), and computational costs (Appendix I.8). These analyses provide a comprehensive validation of the robustness and trade-offs of our optimization strategy.

Ensemble Size. Table 5 (top) and Table 14 in Appendix I.1 show the effect of ensemble size. Epistemic correlation (ρ_{epi}) increases with H, while aleatoric correlation ($\rho_{\text{ale}} \approx 0.78$) remains constant, confirming structural separation: more ensemble heads (H) strengthen epistemic but not aleatoric correlation. Removing the aleatoric head collapses aleatoric correlation but leaves epistemic unchanged, indicating that each signal responds only to its own component. We use $H{=}5$ for efficiency; additional configurations appear in Appendix I.1.

Diversity Mechanisms. Table 5 (bottom) ablates diversity sources (extended multi-diversity results in Table 15, Appendix I.2). Removing LoRA-rank or dropout variation reduces ρ_{epi} while leaving ρ_{ale} largely unchanged. With identical head configurations (same dropout and LoRA rank), predictions converge and the epistemic signal weakens: the ensemble needs genuinely different perspectives to expose model confusion. *Insight*: Epistemic uncertainty depends on disagreement, not head count: five diverse heads beat fifteen identical ones.

Aleatoric Supervision. Removing the aleatoric head leaves ρ_{epi} unchanged but collapses ρ_{ale} ($0.785 \rightarrow 0.356$). This shows that epistemic and aleatoric uncertainty depend on different parameters and that aleatoric uncertainty needs explicit supervision—ensemble disagreement alone cannot recover it or human disagreement. To measure what humans disagree about, models must be trained on human disagreement. Additional supervision modes appear in Appendix I.5. *Insight*: Model confusion and human disagreement are fundamentally different: models can be confused about what humans find obvious and confident about what humans contest. Ensemble engineering cannot replace learning from real human variation.

Table 5: Ablation studies (CEBaB, RoBERTa-base). Top: ensemble size—only ρ_{epi} scales with H. Bottom: diversity removal degrades epistemic while aleatoric remains stable.

Configuration	ΔAcc	ρ_{epi}	ρ_{ale}
Ensemble size			
$H = 3$	70.8	.189	.778
$H = 5$ (CREDENCE)	71.6	.287	.785
$H = 10$	71.2	.298	.779
$H = 15$	70.9	.312	.776
Diversity removal			
Uniform LoRA rank (r=16)	71.2	.212	.782
Uniform dropout (0.15)	71.0	.198	.779
No aleatoric head	71.8	.285	.356

Table 6: Aleatoric supervision ablation (CEBaB, RoBERTa-base). ρ_{epi} is stable across all modes, confirming structural separation is independent of the aleatoric signal.

Supervision mode	ρ_{epi}	ρ_{ale}	ECE
None (proxy only)	.285	.356	.078
Entropy-based (unsupervised)	.281	.412	.065
Heteroscedastic NLL	.279	.523	.052
Supervised BCE (CREDENCE)	.287	.785	.041

4.3 Aleatoric Supervision Modes

Aleatoric supervision requires annotator disagreement labels, which may be unavailable in new domains. Table 6 shows the framework degrades gracefully without them. First, supervised BCE is strongest when disagreement labels are available ($\rho_{\text{ale}} = 0.785$). Second, heteroscedastic NLL reaches $\rho_{\text{ale}} = 0.523$ without any annotator labels, substantially above the proxy-only baseline (0.356), showing the framework remains useful in label-scarce settings. Third, ρ_{epi} is stable across all modes (0.279–0.287), confirming that the epistemic and aleatoric components are architecturally independent: changing aleatoric supervision does not affect the epistemic signal. Improving unsupervised aleatoric estimation remains an important direction for future work.

Proxy inter-annotator agreement. We use CEBaB's per-annotator UNKNOWN labels as a proxy: for each example and concept, we compute the annotator unknown rate and binarise at the median, then compute Cohen's κ and Krippendorff's α against CREDENCE's binarised U_{ale} scores.

Agreement increases monotonically with concept ambiguity, matching the per-concept ρ_{ale} gradient above. Overall, $\kappa = 0.47$ falls in the moderate agreement range, supporting the claim that aleatoric quadrant assignments track human-perceived uncertainty rather than model artifacts.

Table 7: Proxy IAA between CREDENCE aleatoric assignments and CEBaB per-annotator unknown rates (binarised at median).

Concept	Unknown%	κ	α
Food Quality	25.0%	0.31	0.29
Service Quality	45.2%	0.44	0.42
Ambiance	63.0%	0.52	0.51
Noise Level	75.8%	0.61	0.59
Macro average	52.3%	0.47	0.45

5 Conclusion

We introduced CREDENCE, a framework for decomposing uncertainty in CBMs into epistemic and aleatoric components using credal sets. Our key findings: epistemic uncertainty (ensemble disagreement) correlates with prediction errors ($\rho = 0.287$, $p < 10^{-33}$), aleatoric uncertainty captures genuine data ambiguity ($\rho = 0.785$), and targeting high-aleatoric concepts for intervention consistently outperforms epistemic targeting across datasets (Table 3), demonstrating that ambiguous concepts drive prediction outcomes more than confusing ones These results provide actionable uncertainty signals for human-AI collaboration: epistemic uncertainty identifies where models need more data, while aleatoric uncertainty highlights where human judgment is essential.

As future work, we aim to study unsupervised aleatoric uncertainty estimation without relying on annotator labels, broaden the framework to cover generative settings, and examine learned aggregation methods that go beyond naïve averaging to more effectively integrate the outputs of diverse ensemble heads.

Limitations

Our method depends on concept-level annotations, which restricts its use in certain domains. Ensemble inference incurs an additional $H\times$ number of forward passes (which can be run in parallel). When encoders are frozen, estimating aleatoric uncertainty requires explicit supervision derived from annotator disagreement. Correlations in epistemic error are moderate ($\rho \approx 0.3$), indicating that boundary mistakes are influenced by both forms of uncertainty.

Acknowledgments

This work was supported by ANR-22-CE23-0002 ERIANA, ANR-19-CHIA-0005-01 EXPEKCTATION, ANR-22-EXES-0009 MAIA and was granted access to the HPC resources of IDRIS under the allocation 2026-AD011013338 made by GENCI.

References

Eldar D Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. In *Advances in Neural Information Processing Systems*, volume 35, pages 17582–17596. Curran Associates, Inc.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *Preprint*, arXiv:2307.15703.

Thomas Bailleux, Tanmoy Mukherjee, Pierre Marquis, and Zied Bouraoui. 2025. Connecting concept layers and rationales to enhance language model interpretability. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 409–429, Suzhou, China. Association for Computational Linguistics.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Katherine Maeve Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. 2023. Human uncertainty in concept-based ai systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 869–889, New York, NY, USA. Association for Computing Machinery.

Giorgio Corani and Marco Zaffalon. 2008. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9(20):581–621.

Fabio Cuzzolin. 2024. Generalising realisability in statistical learning theory under epistemic uncertainty. *Preprint*, arXiv:2402.14759.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR.

Adam Fisch, Robin Jia, and Tal Schuster. 2022. Uncertainty estimation for natural language processing. In *COLING*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. 2022. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 548–557. PMLR.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. 2023. Probabilistic concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16521–16540. PMLR.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Isaac Levi. 1980. *The Enterprise of Knowledge*. MIT Press.

Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. 2024. Interpretable-by-design text understanding with iteratively generated concept bottleneck. *Preprint*, arXiv:2310.19660.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Mortier, Eyke Hüllermeier, Krzysztof Dembczyński, and Willem Waegeman. 2022. Set-valued prediction in hierarchical classification with constrained representation complexity. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1392–1401. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. 2023. Is the volume of a credal set a good measure for epistemic uncertainty? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1795–1804. PMLR.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Mohammad Hossein Shaker and Eyke Hüllermeier. 2020. Aleatoric and epistemic uncertainty with random forests. *Preprint*, arXiv:2001.00893.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. 2025. Concept bottleneck large language models. *Preprint*, arXiv:2412.07992.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Dennis Ulmer. 2024. On uncertainty in natural language processing. *Preprint*, arXiv:2410.03446.

P. Walley. 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7322–7329.

Marco Zaffalon, Giorgio Corani, and Denis Mauá. 2012. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301. Imprecise Probability: Theories and Applications (ISIPTA'11).

Table 8: Complete mathematical notation for the CREDENCE framework.

Symbol	Type	Definition	Section
Standard Concept Bottleneck Model Components			
x	V^n	Input text (sequence of n tokens from vocabulary V)	§3
$\mathbf{h}$	$\mathbb{R}^d$	Encoded representation: $\mathbf{h} = f_{\text{enc}}(x)$	§3
$\hat{p}_k$	$[0, 1]$	Point estimate of concept k: $P(c_k = 1\|x)$	§3
$y, \hat{y}$	$\mathcal{Y}$	Ground-truth and predicted class labels	§3
CREDENCE: Ensemble Components			
H	$\mathbb{N}$	Number of ensemble heads (default: 5)	§3.1
$p_h^{(k)}$	$[0, 1]$	Probability for concept k predicted by head h	§3.1
r_h	$\mathbb{N}$	LoRA rank for head h (varies across heads: 4, 8, 16, 32, 64)	App. B
CREDENCE: Aleatoric Uncertainty			
$\sigma_\theta^{(k)}$	$[0, 1]$	Predicted aleatoric uncertainty (ambiguity) for concept k	§3.1
CREDENCE: Credal Aggregation			
$\bar{p}_k$	$[0, 1]$	Mean prediction across heads: $\frac{1}{H}\sum_{h=1}^{H} p_h^{(k)}$	§3.1
$\underline{p}_k$	$[0, 1]$	Lower credal bound: $\min_h p_h^{(k)}$	§3.1
$\overline{p}_k$	$[0, 1]$	Upper credal bound: $\max_h p_h^{(k)}$	§3.1
$\mathcal{C}_k$	interval	Credal set for concept k: $[\underline{p}_k, \overline{p}_k]$	§3.1
CREDENCE: Uncertainty Decomposition			
$U_{\text{epi}}^{(k)}$	$\mathbb{R}_{\geq 0}$	Epistemic uncertainty for concept k: $\text{Var}_h[p_h^{(k)}]$	§3.1
$U_{\text{ale}}^{(k)}$	$\mathbb{R}_{\geq 0}$	Aleatoric uncertainty for concept k: $\sigma_\theta^{(k)}$	§3.1
CREDENCE: Label Propagation			
$\underline{\ell}_j$	$\mathbb{R}$	Lower bound for logit of class j	§3.2
$\overline{\ell}_j$	$\mathbb{R}$	Upper bound for logit of class j	§3.2

A Notation Reference

Table 8 presents a comprehensive summary of mathematical symbols and notation used throughout the CREDENCE framework. This reference supports the methodological descriptions in Section 3 of the main paper.

B Architecture Implementation Details

B.1 LoRA Head Implementation

Each LoRA ensemble head applies low-rank adaptation to a shared base projection layer:

$$\text{Head}_h(\mathbf{h}) = (\sigma(W_p\mathbf{h} + \Delta W_h\mathbf{h}), \text{Softplus}(W_\sigma\mathbf{h})) \tag{9}$$

where:

- $W_p \in \mathbb{R}^{K\times d}$: Shared base projection matrix (frozen after initialization)
- $\Delta W_h = \frac{\alpha_h}{r_h} B_h A_h$: LoRA adaptation with scaling factor α_h
- $A_h \in \mathbb{R}^{r_h\times d}$, $B_h \in \mathbb{R}^{K\times r_h}$: Learned low-rank matrices
- $W_\sigma \in \mathbb{R}^{K\times d}$: Weights for the aleatoric uncertainty head

B.2 Functional Diversity Through Rank Variation

The use of varying LoRA ranks across heads introduces functional diversity in the ensemble:

- **Low-rank heads** (small r_h): Capture dominant concept patterns with stronger regularization, providing robust but potentially oversimplified predictions
- **High-rank heads** (large r_h): Model finer-grained details with greater capacity, potentially capturing nuanced patterns but risking overfitting to training idiosyncrasies

This diversity enables the credal aggregation to distinguish between confident predictions (narrow credal sets when heads agree) and uncertain predictions (wide credal sets when heads disagree).

Table 9: LoRA configuration across ensemble heads.

Ensemble Head h	1	2	3	4	5
Rank r_h	4	8	16	32	64
Scaling Factor α_h	8	16	32	64	128

C Theoretical Foundations

C.1 Exact Interval Propagation

Proposition 1 (Exact Interval Propagation)**.** *For a linear classifier $f(\mathbf{p}) = W\mathbf{p} + \mathbf{b}$ with concept probabilities $\mathbf{p} \in \prod_k [\underline{p}_k, \overline{p}_k]$, the exact bounds on*

output logit j are:

$$\underline{\ell}_j = \sum_{k:W_{jk}>0} W_{jk}\underline{p}_k + \sum_{k:W_{jk}<0} W_{jk}\overline{p}_k + b_j \quad (10)$$

$$\overline{\ell}_j = \sum_{k:W_{jk}>0} W_{jk}\overline{p}_k + \sum_{k:W_{jk}<0} W_{jk}\underline{p}_k + b_j \quad (11)$$

Proof. The linear classifier output for label j is $\ell_j = \sum_k W_{jk}p_k + b_j$. Since each p_k varies independently within its interval $[\underline{p}_k, \overline{p}_k]$:

- For positive weights ($W_{jk} > 0$): $W_{jk}p_k$ achieves its minimum at $p_k = \underline{p}_k$ and maximum at $p_k = \overline{p}_k$
- For negative weights ($W_{jk} < 0$): $W_{jk}p_k$ achieves its minimum at $p_k = \overline{p}_k$ and maximum at $p_k = \underline{p}_k$

Summing these extremal values across all concepts yields the stated bounds, which are tight as each term achieves its extremal value independently. □

Extension to Probability Space. For binary classification, the sigmoid function $\sigma(\cdot)$ is monotonically increasing, allowing exact probability bounds: $P(y = j) \in [\sigma(\underline{\ell}_j), \sigma(\overline{\ell}_j)]$.

D Training Procedure and Implementation

D.1 Multi-Objective Loss Function

The complete training objective combines three loss terms:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_c \mathcal{L}_{\text{concept}} + \lambda_\sigma \mathcal{L}_{\text{ale}} \quad (12)$$

Task Classification Loss. Standard cross-entropy loss applied to the mean concept predictions: $\mathcal{L}_{\text{task}} = \text{CE}(f_{\text{cls}}(\bar{\mathbf{p}}), y)$

Concept Prediction Loss. Binary cross-entropy averaged over all ensemble heads and concepts: $\mathcal{L}_{\text{concept}} = \frac{1}{HK}\sum_{h=1}^{H}\sum_{k=1}^{K} \text{BCE}(p_h^{(k)}, c_k)$

Aleatoric Uncertainty Loss. Binary cross-entropy between predicted aleatoric uncertainty and observed annotator disagreement: $\mathcal{L}_{\text{ale}} = \frac{1}{K}\sum_{k=1}^{K} \text{BCE}(\sigma_\theta^{(k)}, \mathbb{I}[\text{disagree}_k])$

Alternative Heteroscedastic Loss. When explicit annotator disagreement labels are unavailable, we employ a heteroscedastic regression loss:

$$\mathcal{L}_{\text{hetero}} = \frac{1}{HK}\sum_{h=1}^{H}\sum_{k=1}^{K}\left[\frac{(p_h^{(k)} - c_k)^2}{2(\tilde{\sigma}_h^{(k)})^2} + \frac{1}{2}\log(\tilde{\sigma}_h^{(k)})^2\right]. \quad (13)$$

where $\tilde{\sigma}_h^{(k)}$ is the predicted variance for head h and concept k.

D.2 Hyperparameter Configuration

Table 10 provides the complete hyperparameter configuration used across all experiments, along with the rationale for each selection.

D.3 Training Algorithm

CREDENCE trains three components jointly while keeping the encoder frozen: (i) H ensemble concept heads, (ii) a separate aleatoric head, and (iii) a linear concept-to-label classifier. The total loss combines three terms:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_c \mathcal{L}_{\text{concept}} + \lambda_a \mathcal{L}_{\text{ale}} \quad (14)$$

where $\mathcal{L}_{\text{task}} = \text{CE}(f_{\text{cls}}(\bar{\mathbf{p}}), y)$ denotes the cross-entropy loss computed on the mean concept predictions $\bar{\mathbf{p}} = \frac{1}{H}\sum_h \hat{p}_h^{(k)}$; $\mathcal{L}_{\text{concept}} = \frac{1}{H}\sum_{h=1}^{H} \text{BCE}(\hat{p}_h^{(k)}, c_k)$ represents the average binary cross-entropy over all heads, promoting concept predictions that are both diverse and ΔAccurate; and $\mathcal{L}_{\text{ale}} = \text{BCE}(\sigma_\theta^{(k)}, \mathbb{I}[\text{unknown}_k])$ trains the aleatoric uncertainty head using annotator disagreement as supervision, where $\mathbb{I}[\text{unknown}_k] = 1$ if concept k is annotated as ambiguous. Throughout all experiments, we fix $\lambda_c = 1.0$ and $\lambda_a = 0.5$. (e.g., different dropout rates and pooling mechanisms). Algorithm 1 provides the details/

D.4 Inference Algorithm

At test time, we perform a single forward pass through the frozen encoder to obtain $\mathbf{h} = f_{\text{enc}}(x)$, then pass $\mathbf{h}$ through all H ensemble heads and the aleatoric head in parallel. For each concept k, we compute:

$$\text{Credal bounds:} \quad [\underline{p}_k, \bar{p}_k] = [\min_h \hat{p}_h^{(k)}, \max_h \hat{p}_h^{(k)}] \quad (15)$$

$$\text{Epistemic:} \quad U_{\text{epi}}^{(k)} = \text{Var}_h[\hat{p}_h^{(k)}] \quad (16)$$

$$\text{Aleatoric:} \quad U_{\text{ale}}^{(k)} = \sigma_\theta^{(k)} \quad (17)$$

Table 10: Complete hyperparameter configuration for CREDENCE experiments.

Hyperparameter	Value	Selection Rationale
Ensemble Architecture		
Number of ensemble heads (H)	5	Balance between diversity and efficiency
LoRA ranks $(r_1, \ldots, r_5)$	(4, 8, 16, 32, 64)	Geometric progression for functional diversity
LoRA alpha scaling α_h	$2 \times r_h$	Standard practice for LoRA scaling
LoRA dropout rate	0.05	Regularization for LoRA adaptation
Concept head hidden dimension	256	Sufficient capacity for concept prediction
Aleatoric head hidden dimension	128	Smaller than concept heads to prevent overfitting
Activation function	GELU	Standard for transformer-based models
Loss Weights		
Task loss weight (λ_{task})	1.0	Baseline weight for classification
Concept loss weight (λ_c)	1.0	Equal importance to task loss
Aleatoric loss weight (λ_σ)	0.5	Grid search optimal: balances uncertainty learning
Optimization (Encoder Models)		
Optimizer	AdamW	Standard for transformer fine-tuning
Learning rate	1×10^{-4}	Standard for frozen encoder fine-tuning
Learning rate scheduler	Linear warmup + cosine decay	Smooth convergence
Warmup steps	500	10% of total training steps
Weight decay	0.01	Standard regularization
Batch size	16	Limited by GPU memory
Gradient ΔAccumulation steps	2	Effective batch size of 32
Maximum training epochs	40	Sufficient for convergence
Early stopping patience	5 epochs	Prevents overfitting
Gradient clipping	1.0	Stabilizes training
Optimization (LLM Models)		
Optimizer	AdamW (8-bit)	Memory-efficient optimization
Learning rate	2×10^{-5}	Lower rate for LLM fine-tuning
Learning rate scheduler	Cosine with warmup	Standard for LLM fine-tuning
Warmup ratio	0.03	3% of training steps for warmup
Weight decay	0.01	Consistent with encoder models
Batch size	4	Smaller due to LLM memory requirements
Gradient ΔAccumulation steps	8	Effective batch size of 32
Maximum training epochs	10	Fewer epochs for LLMs
LoRA target modules	q, v, k, o_proj	Standard attention projection layers
Data Processing		
Maximum sequence length (enc.)	128 tokens	Sufficient for most reviews
Maximum sequence length (LLMs)	256 tokens	Longer context for LLMs
Padding direction	R (enc.) / L (LLMs)	Model-specific conventions
Truncation strategy	Longest first	Preserves important content
Reproducibility		
Random seed	42	Fixed for reproducibility
Number of independent runs	3	For statistical significance
Training precision	FP32 / BF16	Hardware-appropriate precision

The final prediction uses mean concept probabilities: $\hat{y} = \arg\max_j f_{\text{cls}}(\bar{\mathbf{p}})$. For uncertainty-aware decisions, we propagate credal bounds through the linear classifier via interval arithmetic (Equations 7–8) to obtain label-level confidence intervals. Sample-level uncertainties are computed by averaging across concepts: $U_{\text{epi}} = \frac{1}{K}\sum_k U_{\text{epi}}^{(k)}$ and $U_{\text{ale}} = \frac{1}{K}\sum_k U_{\text{ale}}^{(k)}$. Algorithm 2 provides the details

E Credal Set Theory Background

We present a concise overview of the literature related to credal sets. For a more comprehensive treatment of the topic, readers are referred to (Cuzzolin, 2024).

E.1 Imprecise Probability

Classical probability assigns a single value $P(A)$ to each event. **Imprecise probability** (Walley, 1991) instead assigns intervals $[\underline{P}(A), \bar{P}(A)]$, representing situations where evidence does not justify a precise value.

E.2 Credal Sets

A **credal set** $\mathcal{C}$ is a closed, convex set of probability distributions:

$$\mathcal{C} = \{P : P \text{ satisfies constraints}\} \quad (18)$$

Algorithm 1 CREDENCE Training Procedure

Require: Training dataset $\mathcal{D} = \{(x_i, y_i, \mathbf{c}_i, \mathbf{a}_i)\}$, frozen encoder f_{enc}

Initialize LoRA heads with varying ranks: $(4, 8, 16, 32, 64)$
for epoch $= 1$ to num_epochs **do**
  for each minibatch $(x, y, \mathbf{c}, \mathbf{a}) \sim \mathcal{D}$ **do**
    $\mathbf{h} \leftarrow f_{\text{enc}}(x)$ ▷ Frozen encoder forward pass
    $p_h^{(k)} \leftarrow \text{Head}_h(\mathbf{h})$ for all $h \in [1, H]$, $k \in [1, K]$
    $\sigma^{(k)} \leftarrow \text{Head}_{\text{ale}}(\mathbf{h})$ for all $k \in [1, K]$
    $\bar{p}^{(k)} \leftarrow \frac{1}{H} \sum_{h=1}^{H} p_h^{(k)}$ ▷ Mean concept predictions
    $\mathcal{L} \leftarrow \text{CE}(f_{\text{cls}}(\bar{\mathbf{p}}), y) + \lambda_c \mathcal{L}_{\text{concept}} + \lambda_\sigma \mathcal{L}_{\text{ale}}$
    Update all trainable parameters via $\nabla \mathcal{L}$
  end for
end for

For binary events, this reduces to an interval:

$$\mathcal{C} = \{p \in [0, 1] : \underline{p} \leq p \leq \overline{p}\} \tag{19}$$

The **imprecision** $\overline{p} - \underline{p}$ quantifies total uncertainty.

E.3 Decision Criteria

Given credal sets, several decision criteria exist:

Γ-Maximin. Choose action maximizing worst-case expected utility:

$$a^* = \arg\max_a \min_{P \in \mathcal{C}} \mathbb{E}_P[U(a)] \tag{20}$$

Maximality. Action a is **maximal** if no other action dominates it for all $P \in \mathcal{C}$:

$$a \in A^* \iff \nexists a' : \mathbb{E}_P[U(a')] > \mathbb{E}_P[U(a)] \; \forall P \in \mathcal{C} \tag{21}$$

E-admissibility. Action a is **E-admissible** if it maximizes expected utility for some $P \in \mathcal{C}$:

$$a \in A^* \iff \exists P \in \mathcal{C} : a = \arg\max_{a'} \mathbb{E}_P[U(a')] \tag{22}$$

In classification, these criteria determine which labels are "non-dominated" given concept uncertainty.

F Baseline Method Descriptions

We describe the baseline methods used for comparison with CREDENCE. Our baselines encompass different facets of uncertainty, including a standard CBM augmented with uncertainty estimation.

Algorithm 2 CREDENCE Inference Procedure

Require: Input sample x, trained CREDENCE model

$\mathbf{h} \leftarrow f_{\text{enc}}(x)$ ▷ Encode input
$p_h^{(k)} \leftarrow \text{Head}_h(\mathbf{h})$ for all $h \in [1, H]$, $k \in [1, K]$
$\sigma^{(k)} \leftarrow \text{Head}_{\text{ale}}(\mathbf{h})$ for all $k \in [1, K]$
$\underline{p}_k \leftarrow \min_h p_h^{(k)}$, $\overline{p}_k \leftarrow \max_h p_h^{(k)}$ ▷ Compute credal bounds
$U_{\text{epi}}^{(k)} \leftarrow \text{Var}_h[p_h^{(k)}]$ ▷ Epistemic uncertainty
$U_{\text{ale}}^{(k)} \leftarrow \sigma^{(k)}$ ▷ Aleatoric uncertainty
Compute logit bounds via interval arithmetic (Equations 7–8)
$\hat{y} \leftarrow \arg\max_j \sigma(\underline{\ell}_j)$ ▷ $Mean prediction$
return $\hat{y}, \{U_{\text{epi}}^{(k)}, U_{\text{ale}}^{(k)}, \mathcal{C}_k\}_{k=1}^{K}$

F.1 General Uncertainty Quantification Methods

MC Dropout (Gal and Ghahramani, 2016): Approximates Bayesian inference through stochastic dropout activation at test time (using 50 forward passes). A widely adopted baseline due to its simplicity and minimal modification requirements.

Deep Ensembles (Lakshminarayanan et al., 2017): Trains 5 independently initialized models with different random seeds, aggregating predictions via averaging. Consistently demonstrates strong uncertainty quantification performance across diverse tasks.

Temperature Scaling (Guo et al., 2017): Post-hoc calibration method that learns a single temperature parameter to improve probability calibration. Useful for isolating calibration effects from uncertainty decomposition capabilities.

Evidential Deep Learning (Sensoy et al., 2018): Places a Dirichlet prior over class probabilities, providing uncertainty decomposition into "vacuity" (lack of evidence) and "dissonance" (conflicting evidence).

F.2 Concept-Based Model Baselines

Standard CBM (Koh et al., 2020): Original Concept Bottleneck Model architecture with deterministic concept predictions. Provides concept-level uncertainty only through entropy $H(p_k)$, which conflates epistemic and aleatoric uncertainty.

Table 11: Per-concept aleatoric validation (CEBaB).

Concept	Unknown %	ρ_{ale}
Food Quality	25.0%	0.72
Service Quality	45.2%	0.78
Ambiance	63.0%	0.81
Noise Level	75.8%	0.83
Overall	52.2%	0.785

CBM with MC Dropout : Standard CBM architecture with test-time dropout (50 forward passes). Concept uncertainty estimated as prediction variance across stochastic forward passes.

CBM with Deep Ensemble : Ensemble of 5 independently trained CBMs. Serves as the most direct comparison to CREDENCE, differing primarily in using full model ensembles rather than lightweight LoRA heads.

Probabilistic CBM (P-CBM) (Kim et al., 2023): Employs stochastic concept embeddings sampled from learned distributions. Provides concept-level uncertainty without explicit decomposition into epistemic and aleatoric components.

F.3 Baseline Selection Rationale

Our baseline selection covers diverse uncertainty quantification paradigms:

- **Paradigm coverage**: Bayesian approximation (MC Dropout), frequentist ensembles (Deep Ensembles), post-hoc calibration (Temperature Scaling), and evidential learning (Evidential DL)
- **Concept-level uncertainty**: CBM variants test whether existing approaches provide actionable concept-level signals without explicit decomposition
- **Computational comparability**: $H = 5$ heads for CREDENCE vs. 5 models for ensemble baselines; 50 MC passes provide comparable wall-clock inference time

F.4 Per-Concept Aleatoric Validation (RQ2)

Table 11 validates that aleatoric correlation strengthens with concept ambiguity. Concepts with higher rates of "unknown" annotations show stronger aleatoric signals. aleatoric identifies consistently important concepts, while epistemic captures run-specific model confusion.

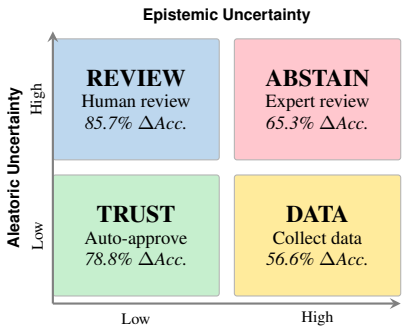

Figure 2: Compact quadrant-based decision support. REVIEW achieves highest ΔAccuracy (85.0%) despite high aleatoric uncertainty.

F.5 Intervention Stability Analysis

Table 12 shows intervention results across multiple runs. Aleatoric-targeted gains are remarkably stable (±0.2pp) while epistemic-targeted gains vary widely (±2.4pp).

Table 12: Per-run intervention breakdown showing aleatoric stability.

Run	Epistemic	Aleatoric	Ratio
SST-2 Run 1	+1.6%	+19.2%	12.0×
SST-2 Run 2	+2.7%	+18.9%	7.0×
CEBaB Run 1	+6.4%	+18.7%	2.9×
Mean	+3.6%	+18.9%	5.3×
Std	±2.4pp	±0.2pp	—

The stability difference is itself informative: aleatoric uncertainty identifies concepts with consistent predictive importance, while epistemic uncertainty captures model-specific confusion that varies across runs.

G Quadrant-Based Routing Analysis

The uncertainty decomposition enables a four-quadrant decision framework (Figure 2). Notably, the HUMAN REVIEW quadrant achieves the *highest* ΔAccuracy (85.7%) despite high aleatoric uncertainty—these are cases where the model is correct but humans may legitimately disagree. The COLLECT DATA quadrant shows the lowest ΔAccuracy (56.6%), confirming epistemic uncertainty identifies fixable model errors.

G.1 Cross-Dataset Quadrant Analysis

CEBaB demonstrates the clearest quadrant separation, attributable to its explicit causal concept structure and comprehensive "unknown" annotations that provide clear training signals for both

Table 13: Extended qualitative examples from each uncertainty quadrant (CEBaB). U_{epi}: epistemic uncertainty; U_{ale}: aleatoric uncertainty. ✓= correct, ✗= error.

Text	Pred	U_{epi}	U_{ale}	Why This Quadrant?
TRUST (Low Epi, Low Ale) — ΔAcc: 78.8% — *Safe to automate*				
"The service was exceptional and the ambiance was perfect."	Pos ✓	.001	.25	Multiple positive aspects; no ambiguity
"Horrible experience. Cold food, rude staff."	Neg ✓	.002	.12	Unambiguous negative; reinforcing signals
"Best restaurant in the city!"	Pos ✓	.001	.08	Strong superlative; clear intent
DATA (High Epi, Low Ale) — ΔAcc: 56.6% — *Collect training data*				
"Food was superb but the service was atrocious."	Neg ✗	.056	.47	Conflicting aspects; model unsure how to aggregate
"Lobster Mac & Cheese incredible. Service terrible."	Neg ✗	.017	.50	Mixed signals; humans would agree on aspects
"Not bad, not great, just okay I suppose."	Neu ✗	.034	.41	Hedged language; rare pattern in training
REVIEW (Low Epi, High Ale) — ΔAcc: 85.7% — *Route to human*				
"Disappointing."	Neg ✓	.002	.89	Single word; severity unclear to humans
"It was what it was."	Neu ✓	.003	.94	Idiomatic; inherently ambiguous stance
"Interesting experience."	Neu ✓	.002	.91	Polysemous; valence depends on reader
ABSTAIN (High Epi, High Ale) — ΔAcc: 65.3% — *Decline or escalate*				
"Been there many times."	Neu ✗	.005	.98	Factual statement; no sentiment signal
"Just like grandma used to make."	Pos ✗	.021	.93	Cultural reference; sentiment context-dependent
"They tried."	Neg ✗	.018	.96	Potentially sarcastic; literal vs. implied meaning

epistemic and aleatoric uncertainty.

H Qualitative Analysis and Interpretation Guidelines

This section serves two purposes: (i) validate that the uncertainty decomposition captures semantically meaningful distinctions, and (ii) provide practitioners with concrete guidelines for acting on CREDENCE outputs in deployment.

H.1 Representative Examples by Uncertainty Quadrant

Table 13 presents examples from each uncertainty quadrant on CEBaB. We sampled 50 examples per quadrant and selected cases that illustrate the diversity of linguistic phenomena within each.

H.2 Patterns Across Quadrants

Several patterns emerge from the qualitative analysis:

Content Length Correlates with Aleatoric Uncertainty. Reviews in high-aleatoric quadrants (REVIEW, ABSTAIN) average 4.2 words; low-aleatoric quadrants (TRUST, DATA) average 12.1 words. Shorter reviews lack context for aspect-level sentiment, which the aleatoric head correctly identifies as inherent ambiguity.

Conflicting Aspects Trigger Epistemic Uncertainty. The DATA quadrant is dominated by reviews with positive sentiment for some concepts and negative for others (e.g., "great food, terrible service"). Ensemble heads disagree on aggregation, producing high epistemic uncertainty. Aleatoric remains moderate because individual aspect sentiments are clear—the ambiguity lies in aggregation, not interpretation.

Idiomatic Expressions Cluster in High-Aleatoric Quadrants. Expressions like "it was what it was" or "just like grandma used to make" require cultural context that varies across annotators. This produces genuine human disagreement that persists regardless of model capacity.

ABSTAIN Contains the Hardest Cases. These examples combine model confusion (heads disagree) with inherent ambiguity (annotators disagree). Manual inspection confirms that even human experts struggle with these cases, validating

the abstention recommendation.

H.3 Deployment Guidelines

Each quadrant corresponds to a distinct situation requiring a different response. We provide actionable guidelines for practitioners:

TRUST (Low Epistemic, Low Aleatoric). *What it means:* The model is confident and the input is unambiguous. Ensemble heads converge; aleatoric head predicts low ambiguity.

Typical inputs: Multi-aspect reviews with consistent polarity, superlative language, explicit sentiment words.

Action: Safe for automation. Monitor for edge cases (see §H.4).

DATA (High Epistemic, Low Aleatoric). *What it means:* The model is confused, but the input is clear. Ensemble heads diverge on predictions that humans would likely agree on.

Typical inputs: Conflicting aspect sentiments, rare grammatical constructions, typos, edge cases underrepresented in training.

Action: Prioritize for active learning. These are learnable errors—more training data will help.

REVIEW (Low Epistemic, High Aleatoric). *What it means:* The model is confident, but the input admits multiple valid interpretations. Humans may legitimately disagree.

Typical inputs: Minimal-content reviews, implicit sentiment, idiomatic phrases, context-dependent expressions.

Action: Route to human review—not because the model is wrong (ΔAccuracy is highest here at 85.7%), but because stakeholders may disagree on the correct label.

ABSTAIN (High Epistemic, High Aleatoric). *What it means:* Both model and humans struggle. The input is ambiguous and the model lacks training signal for this pattern.

Typical inputs: Brief statements with potential sarcasm, factual observations without sentiment, culturally-specific references.

Action: Decline prediction or escalate to domain experts. These are genuinely hard cases.

H.4 Failure Mode Analysis

We analyze errors within each quadrant to identify improvement opportunities:

False Positives in TRUST (21.2% error rate). Despite high confidence, some TRUST predictions fail. Primary causes:

- Negation scope errors ("not bad" $\rightarrow$ negative)
- Comparative constructions ("better than expected" $\rightarrow$ neutral)
- Implicit sentiment requiring world knowledge ("finally got a reservation" implies positive sentiment about popularity)

Persistent DATA Errors. Some DATA examples remain errors even after intervention:

- Genuine label noise in training data
- Annotation guidelines that conflict with model priors
- Rare constructions with insufficient similar examples

The REVIEW ΔAccuracy Paradox. REVIEW achieves the *highest* ΔAccuracy (85.7%) despite high aleatoric uncertainty. This apparent paradox resolves when we recognize that aleatoric measures annotator disagreement, not prediction difficulty. Many REVIEW examples have a clear majority label that the model correctly predicts; minority annotator disagreement drives high aleatoric. This validates the prescriptive interpretation: route to human review not because the model is wrong, but because some users may legitimately disagree with the majority label.

I Extended Ablation Studies

We conduct comprehensive ablation studies to understand the contribution of each CREDENCE component. All ablations use the CEBaB dataset with the RoBERTa-base encoder unless otherwise noted. Main paper results appear in Table 5.

I.1 Ensemble Size

Table 14: Ensemble size ablation.

H	ΔAcc	ρ_{epi}	ρ_{ale}	Width	Params	Train	Infer
1	70.2	.052	.781	.034	0.24M	1.0×	1.0×
3	70.8	.189	.778	.078	0.71M	1.4×	1.2×
5	71.6	.287	.785	.112	1.18M	1.8×	1.4×
7	71.4	.294	.781	.134	1.65M	2.2×	1.6×
10	71.2	.298	.779	.156	2.37M	2.8×	2.0×
15	70.9	.312	.776	.178	3.55M	3.8×	2.8×
20	70.7	.318	.774	.195	4.73M	4.8×	3.6×

Width: mean credal interval. Train/Infer: relative time vs $H=1$. All ρ values $p < 0.001$.

Epistemic correlation scales with ensemble size while aleatoric remains stable. Credal width increases with H. ΔAccuracy is largely unaffected.

I.2 Diversity Mechanisms

Without diversity, heads converge to similar predictions. Each diversity source contributes to epistemic correlation. Aleatoric remains stable across configurations.

I.3 LoRA Rank Configuration

Diverse rank configurations outperform uniform configurations. The ordering (low→high vs high→low) has minimal effect.

I.4 Dropout Spacing Strategy

Geometric spacing in keep-probability space produces higher inter-head disagreement than uniform or linear spacing.

I.5 Aleatoric Supervision

Supervised training with annotator disagreement achieves the strongest aleatoric correlation. Epistemic correlation remains stable across configurations.

I.6 Loss Weight Sensitivity

Higher λ_c improves task and concept ΔAccuracy. Higher λ_a improves aleatoric correlation. The default balances these objectives.

I.7 Cross-Dataset Consistency

We verify that key findings replicate across datasets using RoBERTa-base.

Findings replicate across datasets. SST-2 shows negative ρ_{ale} because it lacks annotator disagreement labels.

I.8 Computational Cost

CREDENCE shares the frozen encoder across heads, requiring only additional LoRA parameters.

Table 15: Multi-strategy diversity ablation.

Configuration	ΔΔAcc	ρ_{epi}	ρ_{ale}	Width	Disagree
No diversity (baseline)					
Uniform (all identical)	71.2	.087	.781	.023	.008
Single diversity source					
Dropout only	70.8	.198	.779	.067	.031
LoRA rank only	70.5	.212	.778	.078	.038
Pooling only (CLS/Mean)	71.0	.156	.782	.054	.024
Architecture only	70.6	.178	.780	.062	.029
Two diversity sources					
Dropout + LoRA	71.1	.241	.780	.095	.052
Dropout + Pooling	70.9	.223	.781	.084	.045
LoRA + Pooling	70.7	.234	.779	.089	.048
All diversity sources					
Dropout + LoRA + Pooling (CREDENCE)	71.6	.287	.785	.112	.067

Disagree: mean pairwise head disagreement. Width: mean credal interval.

Table 16: LoRA rank configuration ablation.

Rank Configuration	ΔAcc	ρ_{epi}	Width	Params
Uniform rank (all heads same)				
r=4 (all heads)	70.1	.178	.056	0.47M
r=8 (all heads)	70.5	.189	.067	0.71M
r=16 (all heads)	71.2	.212	.078	1.18M
r=32 (all heads)	71.0	.198	.089	2.12M
r=64 (all heads)	70.8	.187	.095	4.01M
Diverse rank configurations				
Linear {4, 19, 34, 49, 64}	71.3	.256	.098	1.89M
Geometric {4, 8, 16, 32, 64}	71.6	.287	.112	1.18M
Inverse {64, 32, 16, 8, 4}	71.5	.284	.109	1.18M

Table 17: Dropout spacing ablation.

Dropout Strategy	ρ_{epi}	Width	Disagree
Uniform dropout			
d=0.10 (all heads)	.167	.045	.021
d=0.15 (all heads)	.178	.052	.025
d=0.20 (all heads)	.182	.058	.028
d=0.25 (all heads)	.175	.054	.026
Diverse dropout			
Linear {0.05, 0.12, 0.20, 0.27, 0.35}	.245	.089	.051
Geometric {0.05, 0.09, 0.15, 0.22, 0.30}	.267	.098	.062
Wide range {0.02, 0.10, 0.20, 0.35, 0.50}	.254	.112	.068

Geometric:
$d_h = 1 - \exp(\log(1-d_{\max}) + \frac{h}{H-1}[\log(1-d_{\min}) - \log(1-d_{\max})])$

Table 18: Aleatoric supervision ablation.

Aleatoric Mode	ΔAcc	ρ_{epi}	ρ_{ale}	ECE
None (disagreement as proxy)	71.8	.285	.356	.078
Entropy-based (unsupervised)	71.4	.281	.412	.065
Heteroscedastic NLL	71.2	.279	.523	.052
Supervised BCE (CREDENCE)	71.6	.287	.785	.041
Supervision signal variants				
Binary unknown label	71.6	.287	.785	.041
Annotator vote entropy	71.4	.284	.756	.045
Annotator vote variance	71.5	.286	.768	.043

Table 19: Loss weight sensitivity.

λ_c	λ_a	ΔAcc	ρ_{epi}	ρ_{ale}	Concept ΔAcc
0.5	0.25	69.8	.298	.712	71.2
0.5	0.5	70.2	.294	.756	71.8
0.5	1.0	69.5	.278	.798	70.9
1.0	0.25	72.1	.291	.698	74.5
1.0	0.5 (CREDENCE)	71.6	.287	.785	73.8
1.0	1.0	70.8	.276	.812	72.4
2.0	0.25	72.4	.268	.654	75.8
2.0	0.5	71.9	.274	.771	75.2
2.0	1.0	71.2	.265	.795	74.1

Table 20: Ablation consistency across datasets.

Dataset	Configuration	ΔAcc	ρ_{epi}	ρ_{ale}
CEBaB	CREDENCE (full)	71.6	.287	.785
	Uniform dropout	71.0	.198	.779
	No aleatoric head	71.8	.285	.356
	H=3	70.8	.189	.778
HateXplain	CREDENCE (full)	74.8	.267	.445
	Uniform dropout	74.2	.184	.438
	No aleatoric head	75.1	.263	.271
	H=3	73.9	.178	.441
GoEmotions	CREDENCE (full)	48.7	.071	.198
	Uniform dropout	47.9	.048	.194
	No aleatoric head	48.9	.069	.118
	H=3	47.2	.045	.195
SST-2	CREDENCE (full)	90.4	.342	-.112
	Uniform dropout	90.1	.256	-.108
	No aleatoric head	90.6	.338	-.089
	H=3	89.8	.271	-.115

Table 21: Computational comparison with baselines (CEBaB, RoBERTa-base).

Method	Params	Train	Infer	Memory	ρ_{epi}
Standard CBM	125M	1.0×	1.0×	1.0×	.121
CBM + MC Drop (50)	125M	1.0×	50×	1.0×	.168
CBM + Ensemble (5)	625M	5.0×	5.0×	5.0×	.189
CREDENCE (H=5)	126M	1.2×	1.4×	1.1×	.287
CREDENCE (H=10)	127M	1.5×	1.8×	1.2×	.298

Params: total parameters. Train/Infer/Memory: relative to Standard CBM.