LOW CORRELATION IS NOT SEPARATION: WHAT GRADIENT ISOLATION DELIVERS FOR EPISTEMIC AND ALEATORIC UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Low correlation between epistemic and aleatoric uncertainty scores does not establish that either score measures its intended source. We study this distinction in a concept bottleneck model with separate heads for a prediction-residual target and an annotation-entropy target. We state sufficient conditions for isolation of the heads' direct loss gradients and report a loss-by-parameter audit, while distinguishing this property from the coupling of their trainable inputs. Across five benchmarks with different supervision regimes, the scores exhibit weak association, near-constant outputs, or poor error detection. On CEBaB, both heads carry signal: the residual-based score detects task errors with AUROC 0.72, and the ambiguity score correlates with annotator entropy at 0.22–0.24. Their mutual correlations are -0.14 and 0.00 without a decorrelation penalty. Nevertheless, the model's own predictive confidence has higher mean error-detection AUROC on every benchmark. On CEBaB, removing ambiguity supervision leaves that head nearly constant, while adding the tested decorrelation penalty produces a near-constant head in every seed. These results motivate an evaluation protocol that checks output variation, target alignment, and predictive utility before interpreting low correlation as uncertainty separation.

1 INTRODUCTION

An uncertain prediction can arise because a model lacks relevant evidence or because an input admits several plausible labels. These cases suggest different responses: collecting data may help with the first, whereas the second may require preserving several interpretations. For example, a sentiment model can misclassify a clear restaurant review because it has not learned a rare expression. Annotators can also disagree about an expression such as "interesting food." A single confidence score need not distinguish these situations. When two uncertainty scores are available, their correlation provides one diagnostic of their relationship (Mucsányi et al., 2024). We ask what this diagnostic establishes when the scores receive separate supervision.

Epistemic uncertainty (EU) concerns limitations of the model or available evidence, while aleatoric uncertainty (AU) concerns variation in labels conditional on the input and annotation protocol (Hüllermeier & Waegeman, 2021). Distributional approaches obtain uncertainty measures from predictive distributions or distributions over predictions (Malinin & Gales, 2018; Kuhn et al., 2023), and their components can be strongly correlated (Mucsányi et al., 2024). However, a model's predicted probabilities alone do not identify their discrepancy from an unrestricted true conditional distribution (Tomov et al., 2026). In particular, expected log loss decomposes into the entropy of the true distribution and its KL divergence from the prediction. Both terms depend on that unknown distribution. Changing the output representation alone does not supply the missing information.

We consider a setting with additional supervision. In a supervised latent variable model (SLVM), externally labelled intermediate units lie on the path to the task prediction. Repeated annotations can also provide an ambiguity target for those units. We instantiate this setting with a concept bottleneck model (CBM): one head predicts concepts, a second learns a logit scale from concept-prediction residuals, and a third predicts annotation entropy. The scale admits a credal-set interpretation, but

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

our evaluation concerns the learned scores, not the coverage or utility of the resulting sets. This construction lets us examine what separate supervision contributes and where its interpretation fails.

We distinguish three properties. Gradient isolation concerns which losses directly differentiate through each uncertainty head. Output dependence concerns the relationship between the learned scores on a data distribution. Semantic validity concerns whether the scores track their intended sources of uncertainty. The first property implies neither of the others: separately trained heads can fit dependent targets, and weakly correlated outputs can be uninformative. Moreover, isolation of the head gradients does not imply independent training trajectories when their input projections remain coupled.

Our contribution is an analysis and empirical study of these distinctions. We give sufficient conditions for direct head-gradient isolation and report an audit of the corresponding computation. We then evaluate output variation, alignment with observable proxies, and error detection across five benchmarks. CEBaB (Abraham et al., 2022) supplies per-concept disagreement targets; the other benchmarks use coarser supervision, lack ambiguity labels, or concern instance-level QA diagnostics. Only on CEBaB do both scores carry the evaluated signals, and predictive confidence remains the stronger error detector throughout. The ablations further expose nearly constant heads under missing supervision and the tested decorrelation penalty. These observations support a concrete evaluation protocol; they do not identify a unique cause of failure across datasets or establish an epistemic–aleatoric decomposition.

2 Related Work

Deep ensembles (Lakshminarayanan et al., 2017), MC dropout (Gal & Ghahramani, 2016), and evidential networks (Sensoy et al., 2018) represent uncertainty through distributions over predictions or predictive parameters. Their decompositions need not be functions of the mean prediction alone. Mucsányi et al. (2024) evaluate uncertainty components through target alignment, correctness, and distribution-shift detection, alongside score correlation. Their observation of strong EU–AU correlation on ImageNet motivates examining the learned components, without implying a universal correlation bound. Tomov et al. (2026) show that, for fixed p, unrestricted variation in p^* prevents a function of p alone from identifying $D_{\mathrm{KL}}(p^* \| p)$. Our heads use input representations and external supervision, so the available information differs; the identification limitation remains relevant to their interpretation.

Credal sets represent uncertainty through sets of probability distributions (Walley, 1991; Cozman, 2000). CreINNs (Wang et al., 2025) and Credal Bayesian Deep Learning (Caprio et al., 2023) use such representations in learning. The representation alone does not determine the meaning of its spread, and credal volume need not be an adequate epistemic measure (Sale et al., 2023). We examine a Gaussian logit scale with a set-valued interpretation and a separately supervised ambiguity output. Our results concern the scores; no reported metric evaluates the predictive sets.

CBMs (Koh et al., 2020) expose an intermediate concept layer, and probabilistic CBMs (Kim et al., 2023) represent uncertainty in its predictions. CREDENCE (Mukherjee et al., 2026) already combines ensemble-derived concept intervals with a separately supervised ambiguity output. Here we study a residual-supervised logit scale, the conditions for direct gradient isolation, and diagnostics for uninformative outputs. Separate ambiguity supervision is therefore not itself a distinction from CREDENCE. The absence of a matched comparison also precludes a claim that the scale construction improves on CREDENCE or ordinary auxiliary heads. Appendix B gives further context.

3 Problem Setup

Let $x \in \mathcal{X}$ be an input, $y \in \{1, \dots, J\}$ its task label, and $c = (c_1, \dots, c_C)$ its concept labels, with c_j in a finite label set $\mathcal{Y}_j$. Binary concepts are encoded as $\{0, 1\}$. The model predicts a concept distribution $p_j(\cdot \mid x)$ and a task distribution $p_y(\cdot \mid x)$. For a specified label space and annotation protocol, let $p^*(\cdot \mid x)$ denote the unknown conditional label distribution. We distinguish the ideal quantities

$$U^*_{\mathrm{epi}}(x) = D_{\mathrm{KL}}\big(p^*(\cdot \mid x) \| p(\cdot \mid x)\big), \qquad U^*_{\mathrm{ale}}(x) = H[p^*(\cdot \mid x)] \tag{1}$$

from learned scores U_{epi} and U_{ale}. Here the first quantity is predictive discrepancy, which can include model misspecification; additional data alone need not remove it. Equation (1) fixes the interpretation used in this paper, rather than claiming a unique definition of uncertainty; Appendix A summarizes the notation.

In practice, p^* is unavailable. Repeated annotations yield an empirical distribution $\widehat{p}_{\text{ann}}$, while answer co-occurrence yields a corpus-derived distribution $\widehat{p}_{\text{corp}}$. Their entropies are observable proxies. Annotation entropy can reflect ambiguity, annotation error, and population differences. Corpus entropy can reflect coverage and frequency bias, including underrepresentation of rare valid answers. Neither is an unbiased reference by definition.

Definition 3.1 (Supervised latent variable model)**.** *An SLVM consists of an encoder $h = f_{\text{enc}}(x)$, a latent map $z = g_{\text{lat}}(h)$, and a task predictor $f_{\text{task}}(z)$. The latent units receive external supervision distinct from the task label and lie on the forward path to the task prediction. For the uncertainty construction studied here, the supervised units also require an ambiguity-related target.*

For a CBM, z contains concept predictions. Point annotations supervise their values, and repeated annotations provide per-concept ambiguity targets. For example, three positive and two unknown labels give $(0, 2/5, 3/5)$ in the order (negative, unknown, positive), with entropy approximately 0.673 nats. If only an instance-level ambiguity target is available, an instance-level auxiliary head can be trained, but this alone does not establish the per-unit supervision required by Definition 3.1.

4 Structurally Separated Uncertainty Heads

4.1 Parameterization and predictive sets

The encoder is frozen. Three trainable projections feed the concept and uncertainty heads:

$$h_\mu = W_\mu h, \qquad h_{\text{epi}} = W_{\text{epi}} h, \qquad h_{\text{ale}} = W_{\text{ale}} h. \tag{2}$$

Each concept is binary. For concept j, the outputs are a logit μ_j, an epistemic scale $\sigma_{\text{epi},j} > 0$, and an ambiguity score $\sigma_{\text{ale},j} \in (0, 1)$; across concepts the epistemic covariance is $\Sigma_{\text{epi}} = \text{diag}(\sigma^2_{\text{epi},1}, \ldots, \sigma^2_{\text{epi},C})$. We denote the uncertainty MLP parameters by ϕ_{epi} and ϕ_{ale}, keeping the projection parameters explicit. This distinction matters because a loss can leave the MLPs separate while coupling their trainable inputs. Appendix A gives the head functions, dimensions, and initialization.

For a radius $\tau > 0$, the Gaussian logit parameterization defines the concept-level credal set

$$\mathcal{C}_{\tau,j}(x) = \big[\text{sigmoid}(\mu_j - \tau\sigma_{\text{epi},j}),\ \text{sigmoid}(\mu_j + \tau\sigma_{\text{epi},j})\big]. \tag{3}$$

This interval is the monotone image of $[\mu_j - \tau\sigma_{\text{epi},j}, \mu_j + \tau\sigma_{\text{epi},j}]$ and identifies a convex set of Bernoulli distributions. The radius is not used in training or evaluation. Ten reparameterized draws $\text{sigmoid}(\mu_j + \sigma_{\text{epi},j}\varepsilon)$, with $\varepsilon \sim \mathcal{N}(0, 1)$, are used to form sampled task-probability ranges, but no reported loss or metric uses them; finite sampled extrema are not certified bounds over a credal set. The ambiguity head is trained as an entropy score rather than an observation-noise variance. Likewise, the learned scale is a residual-based score rather than a calibrated estimate of Equation (1).

The task head reads the point concept probabilities $\text{sigmoid}(\mu)$ only; these are not, in general, the means of the sampled sigmoid-transformed Gaussian variables. For strict isolation of the uncertainty heads, its training path must not pass through uncertainty-dependent samples (Appendix C); Figure 1 depicts this condition. The feature regularizer is

$$\mathcal{L}_{\text{orth}} = \|W_\mu W_{\text{epi}}^\top\|_F^2 + \|W_\mu W_{\text{ale}}^\top\|_F^2 + \|W_{\text{epi}} W_{\text{ale}}^\top\|_F^2. \tag{4}$$

It encourages orthogonal row spaces but does not enforce exact orthogonality. Even exact orthogonality would not imply independent features on the data distribution. Moreover, when the projections are trainable, this regularizer couples their updates.

4.2 Supervision and objective

We specify the concept-level losses for CEBaB; the other target mappings are described in Section 5.1 and Appendix K. Let $\mathcal{K}(x)$ be the concepts of x with a positive or negative point label $c_j \in \{0, 1\}$,

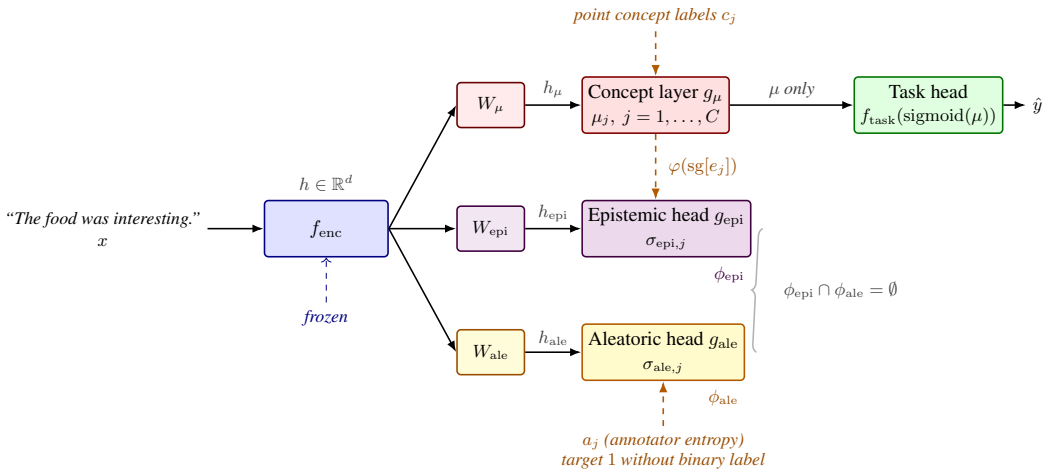

Figure 1: The separate-head construction (CBM instantiation). Solid arrows show the point-prediction path and the inputs to the uncertainty heads; dashed arrows show training signals. The epistemic head is trained against a detached, error-derived target and the ambiguity head against an entropy target; neither target is an inference input. The reported scores are $U_{\text{epi}}(x) = \frac{1}{C}\sum_j \log \sigma_{\text{epi},j}(x)$ and $U_{\text{ale}}(x) = \frac{1}{C}\sum_j \sigma_{\text{ale},j}(x)$. The diagram omits coupling terms for readability: Equation (7) and Proposition 4.1 state when head gradients are isolated. Trainable projections with an orthogonality penalty remain coupled.

and let $a_j = H[\widehat{p}_j]/\log 3$ be the normalized entropy of the annotations of concept j over (negative, unknown, positive). The ambiguity head minimizes

$$\mathcal{L}_{\text{ale}} = \frac{1}{|\mathcal{K}|}\sum_{j\in\mathcal{K}}(\sigma_{\text{ale},j} - a_j)^2. \tag{5}$$

The entropy target is *only* used during training, during inference no such value is provdied. The loss applies only to concepts with a binary point label. Our ambiguity evaluation also includes annotated concepts outside this mask, which instead receive the auxiliary target described below. Thus the reported correlation measures agreement with annotation entropy on a broader set than the entropy-regression training mask; it does not identify latent ambiguity (Appendix E).

The error-related head regresses a detached concept residual:

$$\mathcal{L}_{\text{epi}} = \frac{1}{|\mathcal{K}|}\sum_{j\in\mathcal{K}}\big(\sigma_{\text{epi},j} - \varphi(\text{sg}[e_j])\big)^2, \qquad e_j = |\,\text{sigmoid}(\mu_j) - c_j|, \tag{6}$$

where sg denotes stop-gradient and $\varphi(e) = \text{clip}(\alpha e + \sigma_{\min}, \sigma_{\min}, \sigma_{\max})$ is a clipped affine scaling (Appendix A). Stop-gradient prevents this regression loss from updating the predictor through its target. It does not remove irreducible error from that target. Indeed, for a Bernoulli label with probability q, a correct predictor $p_j = q$ still has $\mathbb{E}[|q - c_j|] = 2q(1-q)$. We therefore describe this head as error-related and assess its epistemic interpretation separately (Appendix E).

Even a correct conditional predictor can yield perfectly correlated residual and entropy proxies while its ideal epistemic divergence is zero (Appendix E).

The complete objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_c\mathcal{L}_{\text{concept}} + \lambda_e\mathcal{L}_{\text{epi}} + \lambda_{\text{KL}}\mathcal{R}_{\text{KL}} \\ + \lambda_a\mathcal{L}_{\text{ale}} + \lambda_u\mathcal{L}_{\text{unk}} + \lambda_o\mathcal{L}_{\text{orth}} + \lambda_d\mathcal{R}_{\text{decorr}}.\end{aligned} \tag{7}$$

Here $\mathcal{L}_{\text{concept}}$ is binary cross-entropy on $\mathcal{K}$, and $\mathcal{R}_{\text{KL}}$ is the concept mean of $D_{\text{KL}}(\mathcal{N}(0, \sigma^2_{\text{epi},j})\|\mathcal{N}(0, \sigma_0^2))$. This regularizer acts on scale only (Appendix A). The auxiliary loss $\mathcal{L}_{\text{unk}}$ regresses $\sigma_{\text{ale},j}$ to 1 on concepts without a binary point label, with $\lambda_u = 0.1\lambda_a$. It therefore associates maximal ambiguity with missing or nonbinary labels, including unannotated aspects.

Finally, $\mathcal{R}_{\text{decorr}}$ penalizes output correlation and couples both heads. The default concept-based models use $\lambda_d = 0$; the CEBaB penalty ablation and QA diagnostics use $\lambda_d = 5$. Appendix A records the configuration and remaining implementation checks.

4.3 What gradient isolation guarantees

Proposition 4.1 (Isolation of uncertainty-head gradients)**.** *Fix a training step. Suppose the encoder is frozen; the two uncertainty MLPs have disjoint parameters; the task and concept losses have no gradient path to either MLP; each supervised uncertainty loss depends only on its own MLP, with detached targets; $\mathcal{R}_{\text{KL}}$ has no path to the ambiguity MLP; and $\mathcal{L}_{\text{unk}}$ has no path to the epistemic MLP. If $\lambda_d = 0$, then*

$$\begin{aligned} \nabla_{\phi_{\text{epi}}} \mathcal{L} &= \lambda_e \nabla_{\phi_{\text{epi}}} \mathcal{L}_{\text{epi}} + \lambda_{\text{KL}} \nabla_{\phi_{\text{epi}}} \mathcal{R}_{\text{KL}}, \\ \nabla_{\phi_{\text{ale}}} \mathcal{L} &= \lambda_a \nabla_{\phi_{\text{ale}}} \mathcal{L}_{\text{ale}} + \lambda_u \nabla_{\phi_{\text{ale}}} \mathcal{L}_{\text{unk}}. \end{aligned} \tag{8}$$

The identities follow by differentiating Equation (7); the orthogonality term contains only projection parameters (Appendix C). They concern direct derivatives at a fixed step. The projections remain coupled when $\lambda_o > 0$, and the detached residual targets change as the concept predictor learns. Consequently, another branch can influence later inputs or targets without a direct cross-head derivative. For $\lambda_d > 0$, the decorrelation derivatives enter both identities. The reported checkpoint audit in Appendix G distinguishes these paths, including those through trainable projections. The checkpoint audit in Table 3 (Appendix G) distinguishes these paths, including those through trainable projections.

Gradient isolation hence establishes a property of the specified computation graph. It does not establish independent training trajectories, low output correlation, or the semantic interpretation of either score. The evaluation below examines these latter properties separately.

5 Experiments

5.1 Evaluation setting

We evaluate direct training dependencies, output variation and association, and alignment with error and ambiguity proxies. The benchmarks differ in their supervision (Tables 7 and 8, Appendix K). CEBaB provides per-concept annotation entropy. HateXplain (Mathew et al., 2022) provides task-vote entropy shared across two units. GoEmotions (Demszky et al., 2020) provides emotion labels but no annotator-entropy target in our construction. MAQA* and AmbigQA*, released by Tomov et al. (2026), provide corpus-derived answer distributions without a concept axis. These QA experiments are instance-level diagnostics outside our SLVM conditions, and use a decorrelation penalty. The primary encoder is frozen DistilBERT; CEBaB also uses RoBERTa-base. Table 1 summarizes three training seeds, $\{42, 123, 2024\}$; exceptions in additional analyses are stated separately.

Data and targets. CEBaB maps star ratings 1–2, 3, and 4–5 to three sentiment classes and uses four concepts: food, service, ambiance, and noise. Its test split contains $1{,}689$ reviews, with majority-class rate 40.3%. Only positive or negative aspect majorities yield binary point labels ($3{,}058$ of $6{,}756$ test slots). The remaining slots enter $\mathcal{L}_{\text{unk}}$; annotation entropy is defined only where annotations exist. Appendix I details this distinction and the earlier naive construction. HateXplain's normalized task-vote entropy has two test values, 0 and 0.579. GoEmotions uses the first listed emotion as a single-label task and 28 latent units; label multiplicity is used only to stratify evaluation. For QA, the ambiguity target is the entropy of corpus answer frequencies and the error-related head outputs one scale per question.

Scores and metrics. For concept-based models, $U_{\text{epi}}(x) = C^{-1} \sum_j \log \sigma_{\text{epi},j}(x) = (2C)^{-1} \log\det \Sigma_{\text{epi}}(x)$. For QA it is the single predicted scale. U_{ale} averages the ambiguity outputs over units. For CEBaB entropy alignment, we instead use $U_{\text{ale}}^{\text{ann}}$, the mean over annotated aspects, and average H over the same mask. This annotation-conditioned diagnostic differs from the all-concept score used for EU–AU correlation; the table abbreviates both as AU. We report task accuracy, Spearman correlations between the two scores and with their evaluation proxies, and task-error detection AUROC. The residual head's training target is concept error, whereas this AUROC concerns task error. MaxProb uses the same model's maximum class probability as a confidence reference; larger

Table 1: Results across five benchmarks, frozen encoders, mean ± SD over seeds $\{42, 123, 2024\}$. "—": withheld under the near-constancy criterion (score SD below 10^{-3}); $(n{=}k)$: reported for k seeds. The two CEBaB ablations use DistilBERT. The penalty ablation produces a near-constant ambiguity head in one seed and a near-constant error-related head in two. QA rows use $\lambda_d = 5$ and are outside the SLVM conditions. AUROC evaluates task-error detection; n.a. denotes an unavailable entropy target.

Benchmark	Configuration	Acc. (%)	ρ EU–AU	ρ EU–Err	ρ AU–H	AUROC U_{epi}	AUROC MaxProb
CEBaB	DistilBERT	77.8 ± 0.1	-0.14 ± 0.03	0.32 ± 0.02	0.22 ± 0.01	0.72 ± 0.02	0.78 ± 0.01
	RoBERTa-base	78.2 ± 0.4	0.00 ± 0.02	0.32 ± 0.02	0.24 ± 0.01	0.72 ± 0.01	0.79 ± 0.01
	No AU supervision	77.4 ± 0.8	—	0.32 ± 0.01	—	0.72 ± 0.01	0.79 ± 0.00
	Decorrelation, $\lambda_d = 5$	77.2 ± 0.4	—	0.30 $(n{=}1)$	0.22 ± 0.01 $(n{=}2)$	0.71 $(n{=}1)$	0.79 ± 0.00
HateXplain	DistilBERT	57.6 ± 0.4	—	0.19 ± 0.02	—	0.61 ± 0.01	0.68 ± 0.01
GoEmotions	DistilBERT	44.6 ± 0.3	—	0.04 ± 0.04	n.a.	0.52 ± 0.02	0.75 ± 0.00
MAQA*	DistilBERT, $\lambda_d = 5$	48.9 ± 1.7	0.01 ± 0.04	0.01 ± 0.02	0.21 ± 0.03	0.50 ± 0.01	0.63 ± 0.01
AmbigQA*	DistilBERT, $\lambda_d = 5$	52.0 ± 0.6	0.05 ± 0.09	-0.07 ± 0.04	0.11 ± 0.07	0.46 ± 0.02	0.65 ± 0.01

uncertainty and lower confidence indicate greater error risk. On CEBaB, we also report an ensemble of the three seed models and MC dropout at rate 0.1 (Appendix K).

A constant score has undefined correlation. We additionally withhold correlations when a score's sample-level standard deviation is below 10^{-3}, an operational near-constancy threshold. This does not make a nonconstant score's correlation mathematically undefined. Since dispersion depends on score units, the threshold is specific to the reported parameterization; Appendix K reports the observed ranges.

Configurations. The default concept-based models use $\lambda_d = 0$. Two CEBaB DistilBERT ablations remove ambiguity supervision ($\lambda_a = \lambda_u = 0$) or add decorrelation ($\lambda_d = 5$); the QA models also use $\lambda_d = 5$. Checkpoints are selected at the earliest epoch attaining maximum validation task accuracy without using uncertainty scores. Entries report the mean ± sample standard deviation over seeds. Bootstrap intervals resample test instances (Appendix I). Hyperparameters are listed in Table 2, and the remaining training details in Appendices A and I.

5.2 Results

Table 1 distinguishes informative scores, near-constant outputs, and scores with poor error detection. These cases require different interpretations even when a raw correlation appears small.

CEBaB: weak association with distinct observable signals. Without a decorrelation penalty, the score correlations are -0.14 ± 0.03 for DistilBERT and 0.00 ± 0.02 for RoBERTa-base. The ambiguity score correlates with annotation entropy at 0.22 and 0.24, but this association varies across concepts: it is strongest for food and service, weaker for ambiance, and close to zero for noise (Table 10). The residual-based score detects task errors with AUROC 0.72 for both encoders, below MaxProb at 0.78 and 0.79. Thus both outputs carry an evaluated signal, without establishing that they isolate the two uncertainty sources. Additional encoders trained with the naive labels yield different associations (Appendix 5), further illustrating that the construction does not fix output correlation.

CEBaB ablations: missing supervision and a decorrelation penalty. Without ambiguity supervision, the ambiguity score is nearly constant in every seed, while task accuracy and error detection remain similar to the default configuration. This ablation removes the ambiguity MLP's direct learning signal; it does not isolate the effect of architecture from that of supervision. With $\lambda_d = 5$, one uncertainty score becomes nearly constant in every seed: the ambiguity score in one and the residual-based score in two. Figure 2 shows this loss of variation within the first 25 epochs, with similar validation accuracy ; Appendix H gives the per-seed timing. The result concerns the tested penalty and weight. Exact Pearson correlation is invariant to nonzero rescaling, so variance reduction alone does not explain it; the implemented treatment of small denominators must be checked before attributing collapse to a specific mechanism (Appendix A).

HateXplain and GoEmotions: nearly constant ambiguity outputs. Across both datasets the ambiguity output is similar for every example, on every seed. Since it barely varies, we do not report

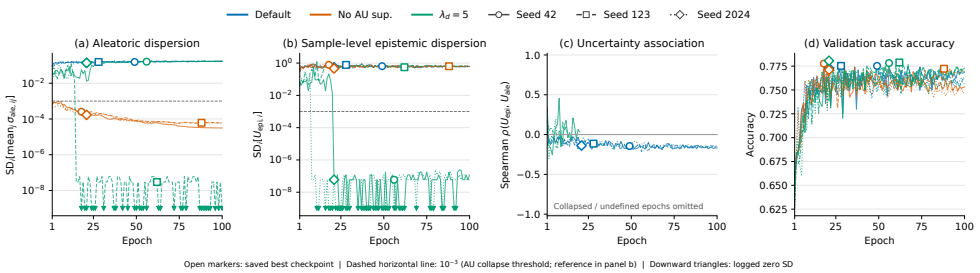

Figure 2: Training dynamics on CEBaB (DistilBERT). (a) Sample-level dispersion of the ambiguity score. (b) Sample-level dispersion of U_{epi}. (c) Validation Spearman correlation of the two scores, omitting epochs with a collapsed head. (d) Validation task accuracy. Open markers: selected checkpoints. Dashed line: the collapse threshold 10^{-3}. Downward triangles: logged zero dispersion.

its correlations: a near-constant score cannot be meaningfully correlated with anything. The output sits right at its training target, but that alone does not tell us which loss or masking choice caused it to saturate. The residual-based score still carries some signal on HateXplain but is close to chance on GoEmotions, and on both datasets it is weaker than simply using the model's confidence. These cases show that a correlation is only meaningful if the score actually varies and is useful in practice.

QA: low association with poor error detection. On MAQA* and AmbigQA*, the score correlations are close to zero (0.01 and 0.05), while error-detection AUROC is 0.50 and 0.46, compared with 0.63 and 0.65 for MaxProb. The ambiguity score tracks corpus entropy weakly (0.21 and 0.11). These diagnostics illustrate that low association can coexist with a poor error score. Because the QA models use $\lambda_d = 5$ and have no concept axis, they do not test the gradient-isolated SLVM setting.

Error detection across ambiguity strata. MaxProb has higher mean AUROC than the residual-based score in every reported stratum (Tables 10 and 11). Its advantage narrows from the lowest to the highest ambiguity group: approximately 0.07–0.08 to 0.05–0.06 on CEBaB, 0.136 to 0.004 on HateXplain, 0.23 to 0.02 on GoEmotions, and 0.15 to 0.05 on MAQA*. These changes concern different proxies, and the MAQA* analysis uses two seeds. MaxProb decreases in all four endpoint comparisons; the residual-based score also changes, including an increase in the small highest-multiplicity GoEmotions group ($n = 63$). The narrowing gap therefore does not by itself establish greater robustness of the residual-based score to ambiguity.

6 Discussion

Our results show that controlling gradient paths is not enough to make two uncertainty scores mean different things: each score must first be shown to vary and learn its own target. On CEBaB, weakly associated scores carry distinct error and ambiguity signals; on the other benchmarks, low correlation arose from near-constant outputs or near-chance error detection (Figure 3).

Gradient isolation makes supervision paths explicit. Proposition 4.1 identifies the losses that directly update each uncertainty MLP. In the no-supervision ablation, the ambiguity MLP therefore loses its direct learning signal. The identities do not determine output dependence. The projections remain coupled, and the residual targets evolve with the concept predictor. Even independently fitted heads would inherit their targets' Spearman correlation if they reproduced the target rankings, including ties (Appendix D). The relevant question is how target dependence and fitting quality relate to output dependence. For dependent targets, forcing decorrelation may conflict with fitting them.

Predicting a proxy does not identify an uncertainty source. The residual head learns concept error, while its reported AUROC concerns task error. These outcomes need not coincide: a task prediction can fail despite correct concepts, and a concept error need not alter the task decision. Moreover, residuals remain positive for ambiguous labels even under a correct conditional predictor (Appendix E). Residual prediction alone therefore establishes neither reducibility with more data nor agreement with the ideal epistemic divergence. MaxProb's higher mean task-error AUROC establishes an advantage for that evaluation objective; it does not make confidence a purely epistemic

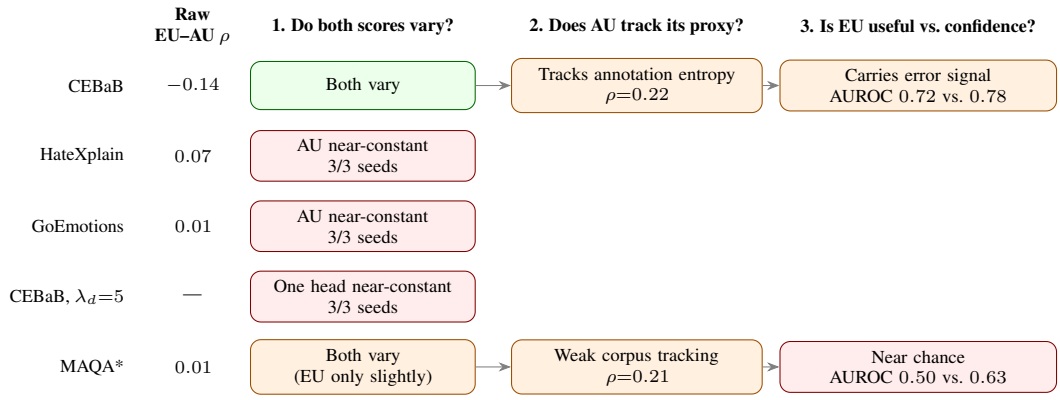

Figure 3: Every configuration shows a low raw EU–AU correlation (left), but only CEBaB passes the protocols that make it interpretable. Checks run left to right and stop at the first failure: score variation (near-constant means sample SD below 10^{-3}), alignment of the ambiguity score with its proxy, and error-detection utility relative to the same model's confidence. Amber marks informative but limited evidence. Values are means over seeds $\{42, 123, 2024\}$ (HateXplain raw ρ: two seeds); the causes of the near-constant heads are given in Section 5.2.The causes for near-constant heads are given in Section 5.2, and per-seed values in Table 6.

quantity or establish that the residual score contains no complementary information. Conversely, weak score association does not demonstrate such complementarity. Whether combining the scores improves decisions remains untested.

The score summary introduces a further distinction. A logit scale and a probability-interval width can rank inputs differently because width also depends on location. In the naive-label runs, replacing mean log scale with mean scale changes AUROC by up to 0.02; replacing a mean with a sum preserves ranks for a fixed concept count (Appendix F). Thus a set-valued representation does not uniquely determine an error-ranking score. The reported aggregation sensitivity does not establish how much of the main gap from confidence is due to score definition or the concept–task mismatch.

The ambiguity score requires the same attention to its target. CEBaB entropy regression uses binary-majority aspects, while the auxiliary objective assigns target 1 to the rest. Entropy evaluation includes annotated aspects from both groups. Table 4 distinguishes $3{,}058$ binary, $1{,}876$ annotated nonbinary, and $1{,}822$ unannotated test slots. Evaluation therefore mixes aspects with different supervision rules. The coarse entropy values also induce many ties. Together with the food–noise contrast in Table 10, this motivates evaluation by concept and supervision mask; it does not establish why a particular concept fails. Training-mask counts and an ablation of the auxiliary loss are needed to distinguish missing supervision from competing targets.

Variation, target alignment, and utility require separate checks. Figure 3 distinguishes near-constant scores from varying scores with near-chance task-error detection. Dispersion must therefore precede correlation analysis, with explicit units, tie handling, and near-constancy criteria. Each score should then be evaluated against its own target before testing downstream utility against confidence and ensemble baselines. Our experiments cover variation and task-error detection, with per-concept ambiguity analysis on CEBaB; direct residual-target alignment, mask-stratified ambiguity analysis, and target correlation are not reported here. Controlled changes in data availability or label ambiguity would further test the proposed uncertainty interpretation. The present evidence supports these diagnostics while leaving semantic separation to be established.

We propose a reporting order for any claimed separation of uncertainty scores: first show that each score varies, with explicit units and a near-constancy criterion; then that each tracks its own target; and only then report their correlation and their utility against confidence and ensemble baselines. Our experiments complete the first and last steps; target alignment and controlled changes in data or label ambiguity remain for future work.

7 Limitations and Conclusion

The positive evidence for both scores rests on CEBaB with two frozen encoders. The other benchmarks vary in tasks, supervision, and, for QA, the objective; their failures do not isolate a single causal mechanism. Ensemble and MC-dropout comparisons are limited to CEBaB. We do not compare against ordinary auxiliary heads with matched targets, vary training-set size, or control annotation noise. Thus we establish neither an advantage of the credal parameterization nor that the residual score measures uncertainty reducible by more data. The ambiguity proxies can reflect annotation bias, label absence, and corpus coverage. Extending the construction to a trainable shared encoder would also require revisiting the gradient paths.

We studied what separate supervision and direct head-gradient isolation establish about uncertainty scores. The reported experiments exhibit weak score association alongside useful proxy signals, near-constant outputs, or poor error detection. Predictive confidence has a higher mean error-detection AUROC on every benchmark. The resulting lesson is operational: verify variation, target alignment, and downstream utility before interpreting low correlation as uncertainty separation. These checks expose failures that gradient identities and correlation alone cannot detect. The direct successive tests will be interventional: if the error-related score is epistemic, it should fall as training data grows, while the ambiguity score should respond to changes in label disagreement but not to data size. We will keep this as future work.

AI Use Statement

We used generative AI tools to gather feedback on writing parts and editing the paper. We also used it for consistency checks on as well as writing code. All experimental work was carried out by the authors, and we accept full responsibility for it.

Reproducibility Statement

Experimental details of training are provided in App A and App I

Ethics Statement

This work uses only publicly available datasets under their released terms; we collected no new data or annotations and ran no studies with human subjects.

References

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior, 2022. URL https://arxiv.org/abs/2205.14140.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal bayesian deep learning. *Trans. Mach. Learn. Res.*, 2024, 2023. URL https://api.semanticscholar.org/CorpusID:257038300.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models, 2023. URL https://arxiv.org/abs/2212.07430.

Fabio G Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions, 2020. URL https://arxiv.org/abs/2005.00547.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. URL https://arxiv.org/abs/1506.02142.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://arxiv.org/abs/1910.09457. arXiv:1910.09457 [cs.LG].

Alireza Javanmardi, David Stutz, and Eyke Hüllermeier. Conformalized credal set predictors, 2024. URL https://arxiv.org/abs/2402.10723.

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, 2022. URL https://arxiv.org/abs/2207.05221.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models, 2023. URL https://arxiv.org/abs/2306.01574.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models, 2020. URL https://arxiv.org/abs/2007.04612.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks, 2018. URL https://arxiv.org/abs/1802.10501.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection, 2022. URL https://arxiv.org/abs/2012.10289.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks, 2024. URL https://arxiv.org/abs/2402.19460.

Tanmoy Mukherjee, Thomas Bailleux, Pierre Marquis, and Zied Bouraoui. Credal concept bottleneck models for epistemic-aleatoric uncertainty decomposition, 2026. URL https://arxiv.org/abs/2604.24170.

Vu-Linh Nguyen, Haifei Zhang, and Sébastien Destercke. Credal ensembling in multi-class classification. *Mach. Learn.*, 114(1), January 2025. ISSN 0885-6125. doi: 10.1007/s10994-024-06703-y. URL https://doi.org/10.1007/s10994-024-06703-y.

Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. Is the volume of a credal set a good measure for epistemic uncertainty?, 2023. URL https://arxiv.org/abs/2306.09586.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty, 2018. URL https://arxiv.org/abs/1806.01768.

Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. The illusion of certainty: Uncertainty quantification for llms fails under ambiguity, 2026. URL https://arxiv.org/abs/2511.04418.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks, 2025. URL https://arxiv.org/abs/2401.05043.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bern Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?, 2023. URL https://arxiv.org/abs/2209.03302.

Marco Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 2002.

A NOTATION AND METHOD DETAILS

We index concepts by $j \in \{1, \dots, C\}$. The task has J labels. The ideal quantities U^*_{epi} and U^*_{ale} in Equation (1) depend on the unknown conditional distribution p^*. The reported scores U_{epi} and U_{ale} are summaries of model outputs. An empirical annotation distribution is denoted by $\widehat{p}_{\text{ann}}$; this notation keeps the evaluation proxy distinct from the underlying population quantity.

Head functions. The encoder output h is the first-token representation of the frozen encoder. Each projection maps h to 256 dimensions; for a 768-dimensional encoder, three 256-dimensional row spaces fit exactly, and the projections are initialized to be mutually orthogonal. The heads are

$$\mu = g_\mu(W_\mu h; \phi_\mu), \tag{9}$$

$$\sigma_{\text{epi}} = \text{clip}\big(\text{softplus}(g_{\text{epi}}(W_{\text{epi}} h; \phi_{\text{epi}})), \sigma_{\min}, \sigma_{\max}\big), \tag{10}$$

$$\sigma_{\text{ale}} = \text{sigmoid}\big(g_{\text{ale}}(W_{\text{ale}} h; \phi_{\text{ale}}) + b\big), \tag{11}$$

where g_μ has one hidden layer of width 128, g_{epi} and g_{ale} have hidden layers of widths 128 and 64, all use GELU activations and dropout 0.1, and b is a learnable per-concept offset initialized at $\text{logit}(0.3)$. The last layer of g_{epi} is initialized so that σ_{epi} starts near the prior scale σ_0. The task head is linear in $\text{sigmoid}(\mu)$.

Error scaling. The target transformation is $\varphi(e) = \text{clip}(\alpha e + \sigma_{\min}, \sigma_{\min}, \sigma_{\max})$ with $\alpha = 2$, $\sigma_{\min} = 0.01$, and $\sigma_{\max} = 2$. Its offset prevents a zero target even when the observed residual is zero, and its clipping limits the target range; neither operation removes ambiguity-related residual error.

Gaussian KL regularization. For $q_j = \mathcal{N}(0, \sigma^2_{\text{epi},j})$ and $q_0 = \mathcal{N}(0, \sigma_0^2)$,

$$D_{\text{KL}}(q_j \| q_0) = \log \frac{\sigma_0}{\sigma_{\text{epi},j}} + \frac{\sigma^2_{\text{epi},j}}{2\sigma_0^2} - \frac{1}{2}. \tag{12}$$

This is the variance part of the Gaussian KL between $\mathcal{N}(\mu_j, \sigma^2_{\text{epi},j})$ and $\mathcal{N}(0, \sigma_0^2)$. The mean does not enter it, so $\mathcal{R}_{\text{KL}}$ has no path to the mean head. The expression compares distributions on the logit scale. It is not equal to a supremum of divergences over the credal set, and it does not give a closed form for a Hausdorff divergence between sets of Bernoulli distributions.

Training configuration. Table 2 lists the default concept-model configuration. The CEBaB ablations remove both ambiguity-supervision terms ($\lambda_a = \lambda_u = 0$) or add decorrelation ($\lambda_d = 5$). The QA runs also use $\lambda_d = 5$, so Proposition 4.1 does not apply to those runs.

Table 2: Default concept-model configuration. The CEBaB ablations and QA runs use the overrides described in the text.

Field	Value	Field	Value
Encoder	frozen	Epochs	100
Projection dimension	256	Prior scale σ_0	0.5
λ_c (concept)	2.0	$\alpha, \sigma_{\min}, \sigma_{\max}$	2, 0.01, 2
λ_e (error)	1.0	Task-probability draws	10
λ_{KL}	0.01	λ_o (orthogonality)	0.001
λ_a (ambiguity)	2.0	λ_d (decorrelation)	0
λ_u (unknown)	$0.1\lambda_a$		

The decorrelation objective. Evaluation uses Spearman correlation. For training, exact ranks are locally constant away from ties, so $\mathcal{R}_{\text{decorr}}$ is specified as the squared batch Pearson correlation of U_{epi} and U_{ale}. It is used in the CEBaB decorrelation ablation and the QA runs. For nonconstant batch vectors u, v and any $a \neq 0$, exact Pearson correlation satisfies $\rho_P(au + b, v)^2 = \rho_P(u, v)^2$. Shrinking one output's variance therefore does not, by itself, reduce this penalty. At zero variance, Pearson correlation is undefined; the numerical behaviour near that point depends on denominator stabilization and any clipping. The observed collapse under $\lambda_d = 5$ does not establish that collapse minimizes the exact Pearson objective.

B Extended related work

Uncertainty methods differ in the information they use as well as in their output representation. An ensemble gives a distribution over predictors, while a separate ambiguity regressor receives additional labels. Comparing their uncertainty correlations therefore requires attention to both supervision and estimation. The critique of conditional entropy and mutual information by Wimmer et al. (2023) further motivates separating a chosen numerical measure from its intended interpretation. Conformal prediction provides a different type of guarantee, concerning coverage under its assumptions (Angelopoulos & Bates, 2022); we do not claim such a guarantee for either uncertainty score or the credal set.

Credal approaches include probabilistic networks (Cozman, 2000), classifiers (Zaffalon, 2002), and neural models (Wang et al., 2025). Other work studies conformalized credal predictors (Javanmardi et al., 2024) and credal ensembling (Nguyen et al., 2025). Our use of a set-valued concept representation does not by itself distinguish all of these approaches. The particular feature examined here is the use of separate supervision and parameter heads for a logit-scale score and an ambiguity score. Whether the set representation improves on an equally supervised auxiliary-head baseline remains an empirical question.

Probabilistic and interactive CBMs study uncertain concepts and concept correction (Kim et al., 2023; Chauhan et al., 2023). CREDENCE (Mukherjee et al., 2026) constructs intervals from diverse concept heads, obtains an epistemic signal from their disagreement, and trains a dedicated ambiguity output against annotation disagreement. Both approaches therefore use separate ambiguity supervision. The present model replaces ensemble-derived spread with a Gaussian logit-scale construction and an error-derived scale target, and analyzes and audits the conditions for gradient isolation. We do not infer semantic validity from that analysis or claim an advantage over CREDENCE without a matched experiment.

For language generation, semantic entropy groups outputs by meaning (Kuhn et al., 2023), while P(True) elicits a model's confidence in an answer (Kadavath et al., 2022). Representation probes instead estimate correctness from internal features (Azaria & Mitchell, 2023). These methods have different inputs and outputs; none should be assigned a pair of EU and AU scores without specifying how that pair is constructed.

C Gradient routing and its limits

Proof of Proposition 4.1. Write ϕ_{epi} and ϕ_{ale} for the parameters of the two uncertainty heads, excluding the projection matrices. The task and concept losses depend on $\text{sigmoid}(\mu)$ only, and therefore on neither head. The target in Equation (6) is detached, and the ambiguity targets are observed label statistics or constants. Hence $\nabla_{\phi_{\text{ale}}}\mathcal{L}_{\text{epi}} = \nabla_{\phi_{\text{ale}}}\mathcal{R}_{\text{KL}} = 0$ and $\nabla_{\phi_{\text{epi}}}\mathcal{L}_{\text{ale}} = \nabla_{\phi_{\text{epi}}}\mathcal{L}_{\text{unk}} = 0$. The orthogonality penalty depends on the projection matrices only. Linearity of differentiation applied to Equation (7) gives

$$\nabla_{\phi_{\text{epi}}}\mathcal{L} = \lambda_e\nabla_{\phi_{\text{epi}}}\mathcal{L}_{\text{epi}} + \lambda_{\text{KL}}\nabla_{\phi_{\text{epi}}}\mathcal{R}_{\text{KL}} + \lambda_d\nabla_{\phi_{\text{epi}}}\mathcal{R}_{\text{decorr}}, \tag{13}$$

$$\nabla_{\phi_{\text{ale}}}\mathcal{L} = \lambda_a\nabla_{\phi_{\text{ale}}}\mathcal{L}_{\text{ale}} + \lambda_u\nabla_{\phi_{\text{ale}}}\mathcal{L}_{\text{unk}} + \lambda_d\nabla_{\phi_{\text{ale}}}\mathcal{R}_{\text{decorr}}. \tag{14}$$

Setting $\lambda_d = 0$ proves the stated identities. The argument concerns direct loss derivatives at a fixed training step. It does not require orthogonal feature projections, establish independence of the learned outputs, or isolate their complete training trajectories.

Trainable projections. Head-level isolation is weaker than isolation of the complete uncertainty branches. For the term $\|W_{\text{epi}}W_{\text{ale}}^\top\|_F^2$ of Equation (4),

$$\nabla_{W_{\text{epi}}}\|W_{\text{epi}}W_{\text{ale}}^\top\|_F^2 = 2W_{\text{epi}}W_{\text{ale}}^\top W_{\text{ale}}, \qquad \nabla_{W_{\text{ale}}}\|W_{\text{epi}}W_{\text{ale}}^\top\|_F^2 = 2W_{\text{ale}}W_{\text{epi}}^\top W_{\text{epi}}, \tag{15}$$

The terms involving W_μ behave analogously. Each derivative depends on another branch's projection. Orthogonal initialization and a finite penalty weight do not ensure exact orthogonality throughout training. When the projections are updated using this penalty, one branch can change the subsequent inputs and gradients of another even though the cross-head loss derivatives remain zero. Fixing the projections or removing the shared regularizer removes this particular path. A shared encoder that is

fine-tuned from both losses, or an optimizer operation such as global gradient clipping, can introduce further coupling.

Detached targets also change across training. In Equation (6), the target at step t depends on the current concept prediction. Stop-gradient blocks differentiation through that target at step t, but an update to the concept predictor changes the target at a later step. Even with fixed projections, the error-related head therefore depends on the concept predictor's training trajectory. The proposition establishes which losses directly differentiate through each uncertainty MLP; it does not assign the learned score exclusively to one supervision source.

The point-prediction path. The specified objective excludes two further gradient paths. Including the mean in the KL regularizer would let that regularizer update the concept predictor. Computing the task loss through uncertainty-dependent concept samples would let task supervision update the scale head. Here, the mean does not enter $\mathcal{R}_{\mathrm{KL}}$, and the sampled probabilities feed only sampled task-probability ranges that carry no loss. Table 3 reports the corresponding gradient pattern for the audited checkpoint.

D Output dependence

Orthogonal projections do not make their outputs statistically independent. For example, let X be uniform on $(0, 1)$, let $h = (X, X)$, and take $W_{\mathrm{epi}} = (1, 0)$ and $W_{\mathrm{ale}} = (0, 1)$. The projection rows are orthogonal, but both projected features equal X. Separate heads can therefore output X and $2X$, giving Spearman correlation one despite orthogonal projections and disjoint parameter blocks. Gradient isolation is neither sufficient nor necessary for zero output correlation; it controls which derivatives reach which parameters.

Accurate fitting can make the outputs inherit the dependence of their targets. Consider a finite evaluation set with no ties in either target. If each head's maximum absolute prediction error is less than half the smallest gap between its target values, each head has the same ranking as its target. Their Spearman correlation then equals the target correlation, independently of whether their gradients are isolated. With ties, preserving the order of distinct target values is not enough: the predictions must also preserve the tie structure. This argument concerns the actual supervised targets on matching units. It does not equate concept-residual alignment with task-error detection, or automatically extend from per-concept targets to aggregated scores.

A decorrelation penalty can conflict with target fitting when the targets are dependent. Its effect also depends on the statistic being optimized: the squared Pearson penalty does not directly minimize the Spearman correlation used in evaluation. Low correlation must therefore be assessed alongside target alignment and error detection. Correlations with zero-variance outputs are undefined. For outputs with small but nonzero variance, correlations can be defined, although they may be withheld under an explicit operational collapse criterion.

E What the supervision targets establish

Prediction residuals are not purely epistemic. The error target in Equation (6) is a prediction-residual proxy. To see its limitation, consider a binary concept with $Y \mid X = x \sim \mathrm{Bernoulli}(q)$ and predicted probability p. Then

$$\mathbb{E}[|p - Y| \mid X = x] = q(1 - p) + (1 - q)p. \tag{16}$$

At $p = q$, this expectation is $2q(1 - q)$, which is positive whenever $0 < q < 1$. In contrast, $D_{\mathrm{KL}}(\mathrm{Bernoulli}(q) \| \mathrm{Bernoulli}(p)) = 0$ at $p = q$. The positive lower bound of φ also prevents the target from vanishing. Residual supervision therefore does not establish that the learned score estimates the ideal divergence or approaches zero with more data. Stop-gradient blocks differentiation through the current target; it removes neither its dependence on the evolving predictor nor its ambiguity-related component.

For a concrete example, let $p(x) = q(x)$ and consider a population squared-loss regressor of the raw residual $|p(x) - Y|$. Its optimal output is $r(x) = 2q(x)(1 - q(x))$. Both $r(x)$ and $H[\mathrm{Bernoulli}(q(x))]$ strictly increase with $q(x) \in (0, 1/2)$. Their Spearman correlation is therefore 1 on any finite set of

at least two inputs with distinct q-values, although the ideal epistemic divergence is zero everywhere. This example concerns unrestricted population regression of the raw residual and entropy; it does not characterize our clipped, masked, and regularized objective.

For squared residuals, the corresponding identity is

$$\mathbb{E}[(p - Y)^2 \mid X = x] = (p - q)^2 + q(1 - q). \tag{17}$$

A correction based on the conditional variance would therefore concern squared loss and require an estimate of $q(1 - q)$. Subtracting an entropy score from an absolute residual has no such identity. Neither correction is part of Equation (6).

Ambiguity supervision is also a proxy. The ambiguity head is trained to predict entropy computed from available annotations. Its agreement with that target measures prediction of the available ambiguity signal, not agreement with an unobserved, unbiased conditional label distribution. Even if annotators were independent draws from a fixed p^*, concavity of entropy gives $\mathbb{E}[H[\widehat{p}]] \leq H[p^*]$ for a finite empirical distribution. Annotator selection and systematic label bias can introduce further discrepancies. These limitations remain under perfect gradient isolation.

F Credal-set geometry

With one logit per concept, $\mathcal{C}_{\tau,j}$ is an interval of Bernoulli parameters, and the convex hull adds nothing. The score $U_{\text{epi}} = \frac{1}{C}\sum_j \log \sigma_{\text{epi},j}$ is rank-equivalent to $\log\det \Sigma_{\text{epi}}$ because the covariance is diagonal and C is fixed. Both are scale statistics in logit space. The width of the probability interval is

$$\text{sigmoid}(\mu_j + \tau\sigma_{\text{epi},j}) - \text{sigmoid}(\mu_j - \tau\sigma_{\text{epi},j}) \approx 2\tau\sigma_{\text{epi},j}\, s_j(1 - s_j), \qquad s_j = \text{sigmoid}(\mu_j), \tag{18}$$

For small $\tau\sigma_{\text{epi},j}$, this approximation shows that interval width depends on both location and scale. For fixed $\tau\sigma_{\text{epi},j}$, the exact width tends to zero as $|\mu_j|$ tends to infinity. A logit-scale score and a probability-scale width can therefore rank inputs differently; neither is calibrated to the ideal epistemic divergence by this construction. In the models trained with the naive label construction (Appendix J), replacing the mean of $\log \sigma_{\text{epi},j}$ with the mean of $\sigma_{\text{epi},j}$ changed AUROC by up to 0.02. Replacing a mean with a sum preserves ranks when the number of concepts is fixed.

G Gradient audit

Table 3 reports a loss-by-parameter audit of one CEBaB DistilBERT checkpoint trained with the naive label construction. The reported pattern agrees with the specified computation graph: task and concept losses use $\text{sigmoid}(\mu)$, samples feed task-probability ranges that carry no loss, the concept-residual target is detached, and the scale KL excludes the mean. At $\lambda_d = 0$, each uncertainty MLP receives gradients only from its own loss terms. The orthogonality penalty couples the trainable projections, and adding $\mathcal{R}_{\text{decorr}}$ introduces gradients in both uncertainty MLPs. The latter configuration is used in the CEBaB decorrelation ablation and the QA runs. This checkpoint audit supports the gradient-routing account for the audited implementation; it does not establish independent training trajectories or verify every reported configuration.

H Training dynamics

Figure 2 shows the validation diagnostics for the three CEBaB DistilBERT configurations and all seeds. With ambiguity supervision, the ambiguity score stays above the collapse threshold throughout training; without it, it stays below the threshold from the first epoch. This control removes both losses that directly train the ambiguity MLP. It therefore tests the behavior of a head without a direct learning signal, while its projection can still change through the orthogonality penalty. The observed lack of variation does not establish that training drives the unsupervised head to a constant. Under $\lambda_d = 5$, the ambiguity score of seed 123 falls below the threshold from epoch 15 onward. Seeds 42 and 2024 retain a varying ambiguity score but lose the recorded variation in the sample-level error-related score. The test outputs show the same pattern at the selected checkpoints.

Table 3: RMS gradient norm of each unweighted loss term with respect to each parameter block (— denotes a reported absent graph path; the encoder has no trainable tensors). The checkpoint is a CEBaB DistilBERT model trained with the naive label construction (Appendix J), seed 123, evaluated over 20 batches in training mode. "Weighted total" uses $\lambda_d = 0$; the final two rows report the decorrelation term and the full objective with $\lambda_d = 5$.

Loss	W_μ	W_{epi}	W_{ale}	mean head	epi head	ale head	task head
Task CE	1.52	—	—	2.73	—	—	0.278
Concept BCE	1.17	—	—	1.72	—	—	—
Error supervision	—	0.267	—	—	0.803	—	—
Scale KL	—	0.923	—	—	3.66	—	—
Ambiguity MSE	—	—	0.336	—	—	0.459	—
Unknown-concept AU	—	—	0.836	—	—	1.43	—
Orthogonality	34.7	19.2	23.3	—	—	—	—
Weighted total	2.91	0.269	0.666	4.68	0.806	0.910	0.278
Decorrelation	—	0.744	1.65	—	1.99	2.13	—
Total, λ_d=5	2.91	3.68	8.34	4.68	9.97	10.7	0.278

Panel (b) measures dispersion across examples after computing U_{epi} for each example. Pooling $\sigma_{\text{epi},j}$ across examples and concepts would answer a different question: differences between concepts can produce nonzero dispersion even when every example receives the same aggregate score. In the default configuration, the correlation between the scores remains about -0.1 to -0.15, and validation accuracy is similar across the three configurations. The selected checkpoints are the earliest epochs of maximum validation accuracy: epochs 49, 28, and 21 for the default; 18, 88, and 21 without ambiguity supervision; and 56, 62, and 21 with $\lambda_d = 5$, for seeds 42, 123, and 2024, respectively.

I Experimental protocol

CEBaB. The task maps star ratings 1–2, 3, and 4–5 to negative, neutral, and positive (test set: $548/461/680$); reviews without a majority rating are dropped. The four concepts are food, service, ambiance, and noise. Table 4 separates annotated and unannotated test aspects and identifies those with a binary target. These test counts describe the evaluation set; they do not establish the frequency or relative contribution of the training losses.

Table 4: CEBaB test aspect slots ($1{,}689$ reviews, $6{,}756$ slots). Every review has at least one annotated aspect. The final column counts annotated aspects without a binary target.

Aspect	Annotated	Unannotated	Binary target	Annotated nonbinary
Food	1,589	100	1,275	314
Service	1,297	392	875	422
Ambiance	1,101	588	546	555
Noise	947	742	362	585
Total	4,934	1,822	3,058	1,876

Naive and corrected label construction. The naive construction maps unannotated aspects to a negative label with entropy 0, and no-majority aspects to a negative label with their observed entropy. It would therefore treat $1{,}822$ unannotated test slots as negative concepts with perfect agreement. The corrected construction assigns neither kind of aspect a binary target. It leaves the entropy of unannotated aspects undefined and excludes them from the review-level entropy H; where annotations exist, both constructions retain the observed entropy. The CEBaB models in Table 1 use the corrected construction. Appendix J reports the separate runs trained with naive targets.

The training and evaluation masks differ. Equation (5) regresses annotation entropy only on the binary-majority set $\mathcal{K}(x)$, whereas ambiguity evaluation includes all annotated aspects. Annotated aspects outside $\mathcal{K}(x)$ instead receive target 1 through $\mathcal{L}_{\text{unk}}$, as do unannotated aspects. Thus an evaluation aspect can have an observed entropy that differs from the target used to train its ambiguity output.

Variation across runs. Unless a caption specifies otherwise, tables report the mean and sample standard deviation over training seeds $\{42, 123, 2024\}$. The naive-label runs and the stratified MAQA* analysis use seeds $\{123, 2024\}$. Entries computed on fewer noncollapsed runs state that number explicitly. Bootstrap intervals resample test reviews and quantify test-set sampling uncertainty, which is distinct from variation across training seeds. Neither type of variation establishes the semantic interpretation of a learned score.

J TRAINING WITH THE NAIVE LABEL CONSTRUCTION

Table 5 reports models trained with the naive label construction and evaluated with the corrected one, using seeds $\{123, 2024\}$ and including two additional encoders. These runs yield different score correlations across encoders. Their error-detection AUROCs for DistilBERT and RoBERTa-base are also lower than those of the corresponding corrected-label configurations in Table 1. Because the training targets and reported seed sets differ, this comparison is descriptive; it does not isolate the effect of label construction in a matched experiment.

Table 5: Models trained with naive labels and evaluated with corrected labels. Mean $\pm$ SD over seeds $\{123, 2024\}$. †Seed 123 only: the seed-2024 ambiguity score has two distinct, nearly equal values over 1,689 reviews and meets the collapse criterion. Its raw correlation is $+0.23$, but is withheld from the aggregate. n.r.: not recomputed.

Encoder	Acc.	Concept acc.	$\rho(U_{\text{epi}}, U_{\text{ale}})$	$\rho(U_{\text{epi}}, \text{Err})$	$\rho(U_{\text{ale}}, H)$	AUROC$_{U_{\text{epi}}}$	AUROC$_{\text{MaxProb}}$
DistilBERT	76.1 ± 0.5	73.6 ± 0.3	-0.09 ± 0.04	0.21 ± 0.06	0.23 ± 0.04	0.64 ± 0.04	0.76 ± 0.01
RoBERTa-base	77.6 ± 0.8	74.5 ± 1.1	-0.03 ± 0.03	0.22 ± 0.02	0.30 ± 0.03	0.66 ± 0.01	0.80 ± 0.00
DeBERTa-v3	76.7 ± 0.1	69.2 ± 0.4	$-0.49^{\dagger}$	n.r.	$0.19^{\dagger}$	0.60 ± 0.00	0.80
ModernBERT	69.7 ± 0.3	66.7 ± 1.8	-0.15 ± 0.02	n.r.	0.20 ± 0.01	0.59 ± 0.01	0.74

Per-seed diagnostics. Table 6 lists the per-seed values summarized in Figure 3.

Table 6: Per-seed values behind Figure 3. SD is the standard deviation of each score across test examples; entries below 10^{-3} meet the near-constancy criterion. "—": correlation withheld because the ambiguity score is near-constant. †Error-related score constant, so its AUROC is 0.5 by construction. n.r.: not included in this export (Table 1 reports these two seeds).

Configuration	Seed	SD(U_{epi})	SD(U_{ale})	$\rho(U_{\text{ale}}, H)$	AUROC U_{epi}	AUROC MaxProb
CEBaB	42	0.630	0.154	0.233	0.732	0.782
	123	0.745	0.152	0.214	0.734	0.769
	2024	0.526	0.140	0.209	0.705	0.775
HateXplain	42	0.210	1.7×10^{-5}	—	0.604	0.671
	123	0.208	6.9×10^{-6}	—	0.622	0.676
	2024	0.231	5.8×10^{-5}	—	0.598	0.698
GoEmotions	42	0.089	1.2×10^{-8}	—	0.513	0.742
	123	0.090	2.4×10^{-9}	—	0.548	0.745
	2024	0.087	1.5×10^{-8}	—	0.507	0.750
CEBaB, λ_d=5	42	0	0.159	n.r.	$0.500^{\dagger}$	0.786
	123	0.557	0	—	0.709	0.787
	2024	0	0.139	n.r.	$0.500^{\dagger}$	0.784
MAQA*	42	0.022	0.049	0.172	0.516	0.643
	123	0.005	0.072	0.225	0.499	0.631
	2024	0.004	0.070	0.223	0.494	0.626

K OTHER BENCHMARKS

Table 7 summarizes the available ambiguity signals and their limitations.

Table 8 records the test-set sizes and target sources.

Table 7: Ambiguity signals and the interpretation of evaluation against them. Where an entropy target is available, it is used for supervision on training instances and evaluated on held-out instances. GoEmotions supplies label multiplicity for stratification, not an annotator-entropy target.

Benchmarks	Ambiguity target	Interpretation and limitation
CEBaB	Per-unit entropy of empirical aspect annotations	Predicts an empirical entropy proxy; may also reflect annotation error, bias, or sparse labels.
HateXplain	Entropy of task-label votes, shared across units	Binary on the test split; no per-unit disagreement target.
GoEmotions	No annotator-entropy target	The processed labels provide multiplicity for stratification only.
MAQA*, AmbigQA*	Instance-level entropy of normalized answer co-occurrence	Tracks corpus answer frequencies; depends on coverage and keyword matching; no per-unit supervision, so outside Definition 3.1.

Table 8: Test-set sizes and target sources.

Dataset	Test	Latent units	Ambiguity source
CEBaB	1,689	4 concepts	Annotator labels
HateXplain	1,924	2 units	Annotator labels
GoEmotions	5,427	28 units	Label multiplicity (strata only)
MAQA*	464	none (instance-level)	Corpus frequencies
AmbigQA* frequencies	421	none (instance-level)	Corpus frequencies

HateXplain and GoEmotions. The ambiguity score meets the collapse criterion on both datasets in every seed. Its outputs are close to 0.99995: the sample-level standard deviation ranges from 6.9×10^{-6} to 5.8×10^{-5} on HateXplain and from 2.4×10^{-9} to 1.5×10^{-8} on GoEmotions. Every concept-level standard deviation is below 10^{-3} when computed in float64. We withhold correlations involving U_{ale} under this operational criterion. The cause is a supervision failure: on HateXplain the entropy-regression loss was inactive because of a field-name mismatch, and GoEmotions has no entropy target, so on both datasets the ambiguity head was trained only by $\mathcal{L}_{\text{unk}}$ toward 1. We therefore draw no conclusion about how well this head tracks annotator disagreement on these datasets. Near-constancy alone does not make correlation mathematically undefined or establish that a ranking is uninformative. Moreover, an absolute dispersion threshold depends on the score's scale.

The HateXplain test entropy takes two values: 974 examples are unanimous and 950 have a 2–1 vote split ($H = 0.579$). Neither processed benchmark supplies a separate entropy target for each latent unit, so the per-concept disagreement analysis used on CEBaB has no direct counterpart here. The error-related score gives $\rho(U_{\text{epi}}, \text{Err}) = 0.19 \pm 0.02$ on HateXplain and 0.04 ± 0.04 on GoEmotions.

QA ambiguity proxies. For MAQA* and AmbigQA*, the ambiguity proxy is an empirical answer distribution obtained by normalizing corpus co-occurrence counts for the valid answers to a question; Wikipedia-derived frequencies provide the target. The resulting entropy is model-independent in the sense that it is computed without the evaluated model's predictions. It can nevertheless underweight rare valid answers and reflect corpus coverage rather than only semantic ambiguity. Correlation between the ambiguity output and this entropy on held-out questions measures generalization to that empirical target, not an independent measurement of latent ambiguity. A variant that supplies the corpus entropy as an input is excluded from the reported comparison, since it would expose the target to the ambiguity predictor.

Instance-level QA models. The model reads only the question and predicts one of a fixed set of answer slots; a prediction is correct when it selects the target slot. The epistemic target is an error proxy minus the ambiguity output, with gradients stopped, and entropy is measured in nats.

Ensemble and MC-dropout comparison. Table 9 gives the reported CEBaB error-detection results for the three-seed ensemble, MC dropout, and the error-related score. The ensemble uses the same three trained seed models as the individual-model analysis. Numerical similarity between AUROCs

does not establish equivalence, and comparing intervals does not replace a paired analysis of their difference.

QA setup. The model reads only the question text. Candidate answers are not encoded as text; each corresponds to a fixed output slot. A prediction is counted as correct when the highest-probability slot is the target slot, not when it matches any valid answer. Entropy is measured in nats and is not normalized. The epistemic target is the error proxy minus the ambiguity output, with gradients stopped; it differs from the concept-level residual used elsewhere. Evaluation uses the 464 MAQA and 421 AmbigQA test examples available for our recorded evaluation outputs.

Table 9: Reported CEBaB error-detection AUROCs. Brackets give bootstrap 95% intervals; MC-dropout entries give the range across seeds. The aggregation underlying the final row requires confirmation as specified above.

Score	DistilBERT	RoBERTa-base
Ensemble MaxProb	0.775 $[0.751, 0.798]$	0.792 $[0.769, 0.814]$
Ensemble predictive entropy	0.764 $[0.740, 0.786]$	0.775 $[0.750, 0.797]$
Ensemble mutual information	0.722 $[0.693, 0.749]$	0.655 $[0.623, 0.683]$
MC-dropout mutual information	0.685–0.733	0.644–0.717
Error-related score U_{epi}	0.720 $[0.694, 0.745]$	0.722 $[0.697, 0.747]$

CEBaB strata and per-concept tracking. Table 10 gives the CEBaB stratified AUROCs and the per-concept ambiguity correlations discussed in Section 5.2. These correlations evaluate annotated aspects, including those excluded from the entropy-regression loss by the binary-majority mask. The weak association for noise therefore warrants a mask-specific analysis before it is attributed to the amount of training supervision.

Table 10: Left: error-detection AUROC by ambiguity stratum, with review entropy averaged over annotated aspects; low is $H = 0$ (unanimous on annotated aspects); positive-entropy reviews are split at the reported threshold $H = 0.288$. Right: per-concept ambiguity tracking $\rho(\sigma_{\text{ale},j}, H_j)$ on annotated test aspects. Mean $\pm$ SD over three seeds.

Encoder	Stratum	n	U_{epi}	MaxProb
DistilBERT	Low	418	0.728 ± 0.015	0.795 ± 0.004
	Medium	686	0.746 ± 0.012	0.793 ± 0.015
	High	585	0.695 ± 0.024	0.744 ± 0.009
RoBERTa-base	Low	418	0.748 ± 0.013	0.824 ± 0.016
	Medium	686	0.728 ± 0.008	0.807 ± 0.021
	High	585	0.698 ± 0.013	0.762 ± 0.008

Aspect	n	DistilBERT	RoBERTa-base
Food	1,589	0.323 ± 0.025	0.336 ± 0.004
Service	1,297	0.259 ± 0.027	0.259 ± 0.005
Ambiance	1,101	0.138 ± 0.012	0.119 ± 0.011
Noise	947	-0.010 ± 0.011	0.014 ± 0.005

Stratified error detection. Table 11 reports error-detection AUROC within the available ambiguity strata. HateXplain and GoEmotions use seeds $\{42, 123, 2024\}$; the MAQA* analysis uses seeds $\{123, 2024\}$. HateXplain has two observed disagreement levels. GoEmotions strata count the annotated emotion labels reconstructed from the concept labels. Label multiplicity is a proxy for ambiguity and does not measure annotator disagreement; moreover, the reported task evaluates the first listed emotion, not a multi-label prediction. The highest stratum contains 61 three-label and 2 four-label examples, so its estimates have substantial variation across seeds.

Stratification rules. Strata are built from test-set entropy, with quantiles computed by linear interpolation. For MAQA, the thresholds are the one-third and two-thirds quantiles, giving low, middle, and high strata; values tied at a boundary go to the upper stratum. For CEBaB, entropy is averaged over a review's annotated aspects, and the two-thirds quantile is computed over all 1689 reviews, including those with zero entropy: low is zero entropy, high is positive entropy at or above the quantile, and middle is in between. HateXplain is split into zero and positive entropy, and GoEmotions by the number of emotion labels (one, two, or three or more). AUROC gives half credit to ties and is undefined for a stratum with no correct or no incorrect predictions. The $\pm$ values are sample standard deviations across seeds.

Table 11: Error-detection AUROC by ambiguity stratum. HateXplain and GoEmotions report mean $\pm$ SD over three seeds; the MAQA* analysis uses two seeds and reports no dispersion. "—" denotes an unreported stratum count

Benchmark	Stratum	n	U_{epi}	MaxProb
HateXplain	Unanimous	974	0.634 ± 0.008	0.770 ± 0.008
	2–1 split	950	0.576 ± 0.021	0.580 ± 0.025
GoEmotions	One label	4,590	0.528 ± 0.024	0.760 ± 0.003
	Two labels	774	0.540 ± 0.063	0.654 ± 0.014
	Three or more	63	0.584 ± 0.032	0.608 ± 0.067
MAQA*	Low tertile	152	0.476 ± 0.015	0.628 ± 0.030
	Middle tertile	157	0.533 ± 0.041	0.564 ± 0.014
	High tertile	155	0.497 ± 0.031	0.547 ± 0.037