KNOW WHAT TO TRUST AND WHAT TO CHECK: CALIBRATION-FREE AUDITING FOR CONCEPT BOTTLENECK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept bottleneck models expose intermediate predictions that can be inspected and corrected. Ensemble disagreement indicates which concepts are uncertain, but does not establish whether this uncertainty affects the final decision. We study this distinction using an input-dependent ellipsoid constructed from concept predictions. For a linear task head, we derive a closed-form margin that certifies prediction invariance throughout the ellipsoid and gives a sufficient certificate after intersection with the probability box. For diagonal geometry, we decompose the squared first-order loss sensitivity into concept contributions, which are the Shapley values of a fixed additive game. We evaluate selective prediction on three datasets and single-concept oracle replacement on CEBaB, separately from optional robust training. Margin-based selective prediction is comparable to confidence and entropy. On CEBaB, however, the three-seed mean lowers accuracy across the evaluated replacement rules. Flagged instances have higher rates of both successful correction and damage to initially correct predictions. These findings distinguish stability under ensemble disagreement from prediction correctness and the benefit of concept correction.

1 INTRODUCTION

Consider a model that predicts the sentiment of a restaurant review through four concepts: food, service, ambiance, and noise. Several concept predictors may agree about the food but disagree about the service. This disagreement matters for the final decision only if changing the service assessment can affect the predicted sentiment. Probing the uncertain concept therefore reveal nothing about the concept that matters most to the decision. Even a consequential correction can turn a correct prediction into an incorrect one. An audit should distinguish these possibilities.

Concept bottleneck models (CBMs) make such verification possible by exposing an intermediate concept representation (Koh et al., 2020). Stochastic predictions and ensembles provide uncertainty estimates for that representation (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017), while probabilistic and stochastic CBMs explicitly model concept uncertainty and dependence (Kim et al., 2023; Vandenhirtz et al., 2024). These estimates describe variation in concept predictions. Its effect on the final decision also depends on the task classifier: the same variance can be consequential along one concept direction and irrelevant along another. This raises two questions. Can the disagreement reach a decision boundary, and which concepts contribute most to the model's local sensitivity?

We address these questions for an ensemble of concept predictors followed by a linear task head. The ensemble mean gives the nominal concept vector, and its dispersion defines an instance-specific ellipsoid. Comparing this region with the classifier's pairwise logit boundaries yields a closed-form margin. A margin greater than one certifies that the predicted class is uniquely optimal throughout the ellipsoid, and hence throughout its intersection with the probability box. The certificate is conditional on the chosen region: it describes prediction stability without assigning a probability that the concepts or the task label are correct.

The same geometry provides a concept-level attribution. With diagonal covariance, each concept contributes its variance multiplied by its squared loss-gradient coordinate to the squared first-order sensitivity penalty. Holding the local gradient fixed, removing uncertainty on a concept axis removes

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.

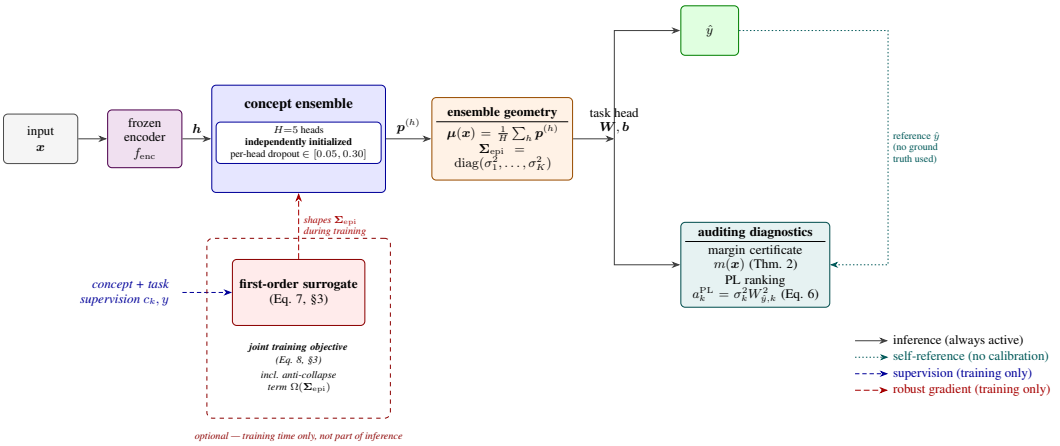

Figure 1: The auditing pipeline. Distinct concept heads share a frozen encoder. Prediction uses the mean concept vector; the margin and ranking use the ensemble geometry and task-head parameters. The loss-gradient ranking additionally requires a specified reference loss and label. Joint robust training is optional and is not part of the inference path.

exactly this contribution. The resulting additive game gives a Shapley interpretation and an optimal selection rule for reducing the fixed penalty. At deployment, we use the predicted class as the reference label. Correcting a concept also changes its mean and possibly the gradient, so the attribution and the realized effect of correction remain separate objects of evaluation.

We further evaluate this through selective prediction and single-concept oracle replacement, and examine how the diagnostics vary with ensemble construction and training. On three datasets, margin-based routing improves over disagreement width and is comparable to confidence and entropy. On CEBaB, the margin identifies a subset with higher rates of both corrected errors and broken correct predictions; every evaluated replacement rule lowers overall accuracy. These results make the scope of the audit concrete: decision sensitivity helps explain which predictions can change, but does not determine whether a change will help. We also compare training configurations, treating joint robust training as an optional way of learning the ensemble rather than a requirement of the certificate.

For a fixed input our uncertain concept set contains all possible values. This allows us to support local robust optimization rather than a distributionally robust guaranty over possible data-generating distributions. The advantage of this analysis combines exact geometry with an explicitly local loss approximation, and the experiments examine what each can establish about model behavior.

2 Model Formulation

We consider a concept bottleneck model with a linear task head. An ensemble estimates the concepts for each input, and its disagreement defines a region of concept vectors (Figure 1). We use this region to assess prediction invariance and attribute local sensitivity. Table 5 collects the notation.

Concept predictions and task head. A frozen encoder maps input $\mathbf{x}$ to $\mathbf{h} = f_{\text{enc}}(\mathbf{x})$. Each of H separately parameterized concept heads produces $\mathbf{p}^{(h)} \in [0,1]^K$, with mean $\boldsymbol{\mu} = H^{-1}\sum_h \mathbf{p}^{(h)}$. The default uses five independently initialized MLPs with training dropout rates between 0.05 and 0.30. The heads share their representation and are not independent statistical samples. Let $\widehat{\sigma}_k^2$ be the across-head sample variance, with divisor $H-1$, and let σ_k^2 denote the effective variance after the recorded standard-deviation clamp (Appendix A). The diagonal construction uses $\boldsymbol{\Sigma}_{\text{epi}} = \text{diag}(\sigma_1^2, \ldots, \sigma_K^2)$. The geometric results allow any positive semidefinite shape matrix, including full and singular matrices. The task head has logits $\mathbf{z}(\mathbf{p}) = \mathbf{W}\mathbf{p} + \mathbf{b}$ and probabilities $\mathbf{q}(\mathbf{p}) = \text{softmax}(\mathbf{z}(\mathbf{p}))$.

An ensemble-disagreement region. We distinguish the ellipsoid from its intersection with a probability box

$$\widetilde{\mathcal{C}} = \{\boldsymbol{\mu} + \boldsymbol{\Sigma}_{\text{epi}}^{1/2}\mathbf{u} : \|\mathbf{u}\|_2 \le 1\}, \qquad \mathcal{C} = \widetilde{\mathcal{C}} \cap [0,1]^K. \tag{1}$$

The square-root definition permits singular matrices: displacement is restricted to the range of $\mathbf{\Sigma}_{\text{epi}}$. For positive definite matrices, $\widetilde{\mathcal{C}}$ is the unit Mahalanobis ellipsoid; in the diagonal case, its semi-axis lengths are σ_k. We summarize its scale by $\epsilon(\mathbf{x}) = \sqrt{\text{Tr}(\mathbf{\Sigma}_{\text{epi}}(\mathbf{x}))}$. The subscript "epi" records the intended interpretation of ensemble disagreement. Neither this interpretation nor the unit-radius choice provides coverage of the true concepts. Appendix B discusses alternative geometries.

Local robust loss For a labelled example, let $\ell(\mathbf{p}) = \text{CE}(\mathbf{W}\mathbf{p} + \mathbf{b}, y)$. The worst-case concept-space loss is

$$\mathcal{L}_{\text{robust}}(\mathbf{x}, y) = \max_{\mathbf{p} \in \mathcal{C}(\mathbf{x})} \ell(\mathbf{p}). \tag{2}$$

This objective varies concept vectors for a fixed input. It is not distributionally robust optimization over data-generating distributions. Its first-order approximation has the following standard support-function form.

Theorem 1 (First-order ellipsoidal robust loss)**.** *Let* $\mathbf{g} = \nabla_{\mathbf{p}} \ell(\boldsymbol{\mu})$ *and* $s = \sqrt{\mathbf{g}^\top \mathbf{\Sigma}_{\text{epi}} \mathbf{g}}$*. Then*

$$\max_{\mathbf{p} \in \widetilde{\mathcal{C}}} \{\ell(\boldsymbol{\mu}) + \mathbf{g}^\top (\mathbf{p} - \boldsymbol{\mu})\} = \ell(\boldsymbol{\mu}) + s. \tag{3}$$

For $s > 0$*, the maximizing vector is* $\mathbf{p}^{\text{lin}} = \boldsymbol{\mu} + \mathbf{\Sigma}_{\text{epi}} \mathbf{g}/s$*. The same value is attained over* $\mathcal{C}$ *if* $\mathbf{p}^{\text{lin}} \in [0, 1]^K$*; otherwise it is an upper bound on the boxed* linearized *objective. For* $s = 0$*, the linearized objective is constant on both sets.*

This identity solves the linearized problem, rather than maximizing cross-entropy itself. Notice how training minimizes the loss over the model parameters, but the inner problem maximizes the loss over the ellipsoid. Appendix C provides the proof, a Taylor-remainder bound, and the exact clipped support calculation.

A certificate of prediction invariance. For a linear task head, the same support calculation determines exactly whether the unboxed region reaches a pairwise logit tie.

Definition 1 (Mahalanobis margin)**.** *Let* $\hat{y} = \arg\max_c z_c(\boldsymbol{\mu})$*, using a fixed tie-breaking rule. For each* $c \neq \hat{y}$*, define* $\mathbf{w}_c = \mathbf{W}_{\hat{y}} - \mathbf{W}_c$*,* $\Delta b_c = b_{\hat{y}} - b_c$*, the nominal gap* $d_c = \mathbf{w}_c^\top \boldsymbol{\mu} + \Delta b_c \geq 0$*, and* $s_c = \sqrt{\mathbf{w}_c^\top \mathbf{\Sigma}_{\text{epi}} \mathbf{w}_c}$*. Set*

$$m_c = \begin{cases} d_c / s_c, & s_c > 0, \\ +\infty, & s_c = 0,\ d_c > 0, \\ 0, & s_c = 0,\ d_c = 0, \end{cases} \qquad m = \min_{c \neq \hat{y}} m_c. \tag{4}$$

Theorem 2 (Prediction-invariance certificate)**.** *The unboxed ellipsoid* $\widetilde{\mathcal{C}}$ *intersects the pairwise tie hyperplane* $\{\mathbf{p} : z_{\hat{y}}(\mathbf{p}) = z_c(\mathbf{p})\}$ *if and only if* $m_c \leq 1$*. Moreover,* $\hat{y}$ *is the unique highest-logit class at every point of* $\widetilde{\mathcal{C}}$ *if and only if* $m > 1$*. Since* $\mathcal{C} \subseteq \widetilde{\mathcal{C}}$*,* $m > 1$ *is also sufficient for this property on* $\mathcal{C}$*; the converse need not hold for the boxed set.*

At $m = 1$, touching a tie need not change the class selected by the tie-breaking rule. We therefore call instances with $m \leq 1$ *not certified*, rather than necessarily unstable. The certificate concerns the chosen region: it establishes neither prediction correctness nor invariance under an oracle correction outside that region. Appendix D gives the proof and exact boxed criterion; Figure 2 illustrates why the boxed converse fails.

2.1 Attributing Local Sensitivity to Concepts

The loss gradient specifies which directions of disagreement matter locally. At inference time, we use cross-entropy with the reference label $\hat{y}$, held fixed at the nominal prediction when differentiating, so $\mathbf{g}_{\hat{y}} = \mathbf{W}^\top (\mathbf{q}(\boldsymbol{\mu}) - \mathbf{e}_{\hat{y}})$. A gold-label gradient is available only retrospectively. For either specified reference loss, diagonal covariance gives

$$\mathbf{g}^\top \mathbf{\Sigma}_{\text{epi}} \mathbf{g} = \sum_{k=1}^{K} \sigma_k^2 g_k^2, \qquad g_k = \frac{\partial \ell}{\partial p_k}(\boldsymbol{\mu}). \tag{5}$$

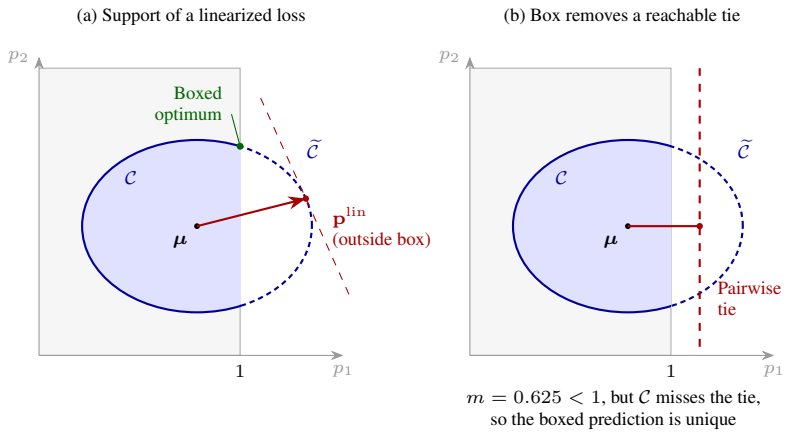

Figure 2: Illustrative geometry. Left: the linearized-loss maximizer $\mathbf{p}^{\text{lin}}$ lies outside the probability box, so the boxed optimum differs. Right: the ellipsoid reaches a pairwise tie that its boxed intersection cannot reach. Hence $m > 1$ is sufficient, but not necessary, for unique prediction invariance on the boxed set.

Thus $a_k = \sigma_k^2 g_k^2$ attributes the squared first-order penalty to concept k. To interpret an inspection, fix $\boldsymbol{\mu}$, $\mathbf{W}$, and $\mathbf{g}$, and remove uncertainty on the inspected axes. The resulting additive game has the following Shapley attribution (Shapley, 1953).

Proposition 1 (Shapley values for a fixed local game)**.** *Let $a_k = \sigma_k^2 g_k^2 \geq 0$ and $A = \sum_j a_j$. For a coalition $S \subseteq [K]$, let $v(S) = \sum_{k \in S} a_k$ be the reduction in the squared penalty A obtained by setting $\sigma_k^2 = 0$ for all $k \in S$, with all other local quantities fixed. Because v is additive, the Shapley value of concept k is $\phi_k = a_k$, and any $k^\star \in \arg\max_k a_k$ gives the largest single-concept reduction. The same $k^\star$ also gives the largest single-concept reduction $\sqrt{A} - \sqrt{A - a_k}$ in the unsquared penalty, although that game is not additive and its Shapley values differ from a_k.*

The proof is in Appendix E. This game attributes the unboxed local penalty, including when the region is clipped. Replacing μ_k with an observed value also changes the nominal loss and usually the gradient; therefore, the proposition provides no optimality guaranty for post-correction accuracy.

Comparison with predicted-logit weighting. We also evaluate variance ranking and a predicted-logit (PL) comparator:

$$a_k^{\text{PL}} = \sigma_k^2 W_{\hat{y},k}^2, \qquad k_{\text{PL}} \in \arg\max_k a_k^{\text{PL}}. \tag{6}$$

PL and gradient score measure different things. PL asks how much each concept moves the score of the predicted class while the gradient score asks how much each concept moves the loss, which depends on how the predicted class compares to other classes. While they disagree in general, even when the model is confident: the gradient is then driven by the difference between the predicted class and the runner-up, not by the predicted class alone. PL also changes if the same vector is added to every class's weights, which leaves the model's predictions unchanged; the gradient score does not (Appendix F). Which ranking corrects more errors is therefore an empirical question.

3 Learning the Ensemble and Its Geometry

The diagnostics apply to a fixed ensemble. To examine how training affects its geometry, we compare three configurations. Mode A uses task and concept supervision and computes covariance after training. Mode B adds a fixed-radius robust term that updates only the task classifier. Mode C jointly updates the ensemble and classifier using an instance-dependent surrogate. We retain *Adaptive-DRO* as the experimental name of Mode C. A two-stage variant starts from Mode A and changes several training settings (Appendix I).

For a labeled example, write $\ell_y(\mathbf{p}) = \mathrm{CE}(\mathbf{W}\mathbf{p}+\mathbf{b}, y)$ and $\mathbf{g}_y = \nabla_{\mathbf{p}}\ell_y(\boldsymbol{\mu})$. The first-order surrogate is

$$\mathcal{L}_{\mathrm{FO}}(\mathbf{x}, y) = \ell_y(\boldsymbol{\mu}) + \sqrt{\mathbf{g}_y^\top \boldsymbol{\Sigma}_{\mathrm{epi}} \mathbf{g}_y}. \tag{7}$$

Its numerical value is the unboxed linearized maximum of Theorem 1. The reported scalar training loss combines it with task and concept supervision:

$$\mathcal{J}(\theta) = \mathbb{E}_{(\mathbf{x},y,\mathbf{c})}\left[\ell_y(\boldsymbol{\mu}) + \lambda_{\mathrm{dro}}\mathcal{L}_{\mathrm{FO}} + \lambda_c\mathcal{L}_{\mathrm{concept}} + \beta\Omega(\boldsymbol{\Sigma}_{\mathrm{epi}})\right], \tag{8}$$

where θ contains the concept-head and classifier parameters, and $\Omega = -\sum_k \log \sigma_k^2$. Numerically, nominal cross-entropy enters with weight $1 + \lambda_{\mathrm{dro}}$. The implementation uses a partial differentiation convention: it forms $\mathbf{g}_y$ at a detached $\boldsymbol{\mu}$ while retaining its classifier dependencies. Consequently, the robust penalty does not propagate derivatives through the center used to form $\mathbf{g}_y$. Appendix G.1 records this convention and the remaining implementation checks.

At fixed gradient, minimizing the penalty favors lower variance in sensitive directions compared to the log barrier opposes collapse with higher variance. Concept supervision and shared parameters also shape the resulting region. Appendix G.2 derives an equalization condition for freely adjustable positive variances; it is not a stationarity condition for the neural ensemble.

4 Experiments

In our experiments we evaluate whether the margin identifies reliable predictions, whether the concept rankings lead to successful corrections, and how training changes these diagnostics. Selective prediction retains a subset of instances and measures its accuracy. Oracle replacement instead changes a concept value and recomputes the prediction. These evaluations distinguish a useful reliability score from a useful intervention rule.

Datasets and protocol. We use CEBaB restaurant reviews with four aspect concepts and a frozen RoBERTa-base encoder (Abraham et al., 2022; Liu et al., 2019); CUB-200-2011 with 112 attributes and frozen DINOv2-base features (Wah et al., 2011; Oquab et al., 2024); and GoEmotions with 28 concept coordinates (Demszky et al., 2020). The GoEmotions task label is the first positive emotion, hence a deterministic function of its annotated concepts. We use this dataset for accuracy and routing. Oracle replacement is evaluated only on CEBaB: the available CUB targets lack the observation mask needed for replacement in probability coordinates.

For CEBaB replacement, a negative annotation becomes 0 and a positive one becomes 1, matching the concept head's output, the probability that the aspect is positive. Aspects marked unknown or left unannotated are never selected. Examples with no eligible aspect keep their original prediction but still count in the accuracy over all test examples. Training treats unknown labels differently, giving them half weight in the loss. To make sure this evaluation uses the same data as training, we re-encoded the raw test records with the original settings; this reproduces the stored labels exactly and the stored features up to numerical precision.

Comparisons and measurement. We compare Modes A–C with CBM and probabilistic concept-head implementations, each using Monte Carlo dropout (50 passes) or a deep ensemble (five models). P-CBM denotes the implementation used in these experiments, whose correspondence to Kim et al. (2023) has not been established. Table 1 gives task accuracy; Appendix L details the available baseline information. Seed counts accompany the results, and $\pm$ denotes standard deviation across seeds. Several comparisons use a single seed.

The baseline geometry also helps interpret these comparisons (Table 9, Appendix L). On CEBaB, CBM-MCD and CBM-DE have almost identical accuracy (.694 and .692), although their widths differ by more than a factor of two (.093 and .216). Disagreement magnitude is therefore not interchangeable with task error across fitted models. Our routing comparison instead evaluates alternative scores on the same checkpoints.

For routing, we compare the margin with disagreement width $\epsilon = \sqrt{\mathrm{Tr}\,\boldsymbol{\Sigma}_{\mathrm{epi}}}$, predictive entropy, and maximum softmax confidence. Mean retained accuracy (MRA) averages accuracy at coverages $\{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0\}$. This is a discrete average, rather than an area under a curve. Retained counts use floor rounding, and ties preserve input order (Appendix M.1). We also report mean width $\bar{\epsilon}$ and Spearman margin–width correlation as descriptive properties of the ensemble.

Table 1: Reported test accuracy for all baseline families. MCD and DE baseline cells use one seed. Mode cells with ± report three seeds. Bold marks the largest reported mean per dataset, without a significance claim.

Method	CEBaB	CUB	GoEmotions
CBM-MCD	.694	.888	.444
CBM-DE	.692	**.896**	.450
P-CBM-MCD	.751	.881	.468
P-CBM-DE	.755	.884	**.473**
Mode A	$.789 \pm .006$	$.857 \pm .001$	.446
Mode B	$.788 \pm .005$	$.809 \pm .002$	.446
Mode C	$\mathbf{.796 \pm .004}$	.817	.445
Mode C, two-stage	—	$.878 \pm .002$	—

Table 2: Mean retained accuracy (MRA) over the seven specified coverage levels. CUB and GoEmotions: single checkpoints. †CEBaB: mean over three seeds (sample standard deviations: margin .006, width .016, entropy .003, confidence .003).

Dataset	Base accuracy	Margin	Width	Entropy	Confidence
CEBaB†	.796	.857	.804	.863	.865
CUB	.878	.956	.896	.947	.956
GoEmotions	.445	.509	.403	.517	.525

Selective prediction. The margin gives MRA close to entropy and confidence on all three datasets (Table 2). It exceeds width-based routing by .053, .060, and .106 on CEBaB, CUB, and GoEmotions, respectively, but does not outperform confidence. Its distinct property is the sufficient stability condition at $m > 1$; the desired retained risk or review budget still requires empirical evaluation.

At threshold $m = 1$, despite the flagged set being small holds a disproportionate share of the errors on both datasets (Table 3). On CEBaB it catches most of the errors. On two-stage CUB it flags only a few instances, and most errors remain among the predictions the certificate marks as stable. A prediction can therefore be stable across the whole ellipsoid and still be wrong. The two CUB rows come from different trained models and should not be compared directly.

The CUB partition also show that a fixed threshold can land at very different operating points. The more accurate two-stage model keeps most of its predictions but is less accurate on those it retains; the baseline keeps fewer and is more accurate on them. Because the threshold keeps different fractions on different models, these numbers do not indicate which model ranks its predictions better. For that, we compare scores at the same coverage levels (Table 2).

We also verify whether the certificate still holds once concepts are kept inside $[0, 1]$ (Table 6, Appendix D). For nearly all flagged cases, the closest point at which the prediction becomes a tie is located within the box. So when the margin indicates the prediction could change, it can also change using valid concept probabilities. Across all audited runs, an exact check with the box almost never certifies an instance that the simple margin does not. For nearly all CEBaB instances, the worst point of the linearized loss also lies inside the box, so the closed-form value is exact there. These checks concern the geometry of the linearized problem; they say nothing about how well it approximates the true cross-entropy.

Single-concept replacement. We evaluate six rules on the original CEBaB Mode C seeds: variance, predicted-label gradient weighting, its retrospective gold-label counterpart, PL, weight-only ranking $\arg\max_k \max_c |W_{c,k}|$, and uniform random selection. Each rule selects among eligible aspects before their values are substituted; ties select the first eligible aspect. Table 4 counts both errors corrected and correct predictions broken, so its accuracy changes describe application to the whole test population.

Every replacement rule lowers accuracy on every seed. Each rule fixes some errors but breaks more predictions that were already correct. Random selection does about as well as the ranking rules, and

Table 3: Partition at $m = 1$ for separate seed-0 runs. Retained: $m > 1$. Brackets: 95% Wilson intervals. Error capture is the fraction of all errors in the flagged set. The missing CUB example identifiers preclude a verified comparison of individual predictions across runs.

Run	N	Errors	Coverage	Retained acc.	Flagged	Error capture
CEBaB, Mode C	1689	343	.707	.885 [.866,.902]	495	.601 [.548,.651]
CUB, Mode C baseline	5794	1062	.690	.953 [.946,.959]	1797	.823 [.799,.845]
CUB, Mode C two-stage	5794	707	.941	.909 [.901,.916]	339	.297 [.265,.332]

Table 4: Single-concept oracle replacement on CEBaB, Mode C, seeds 0/1/2. Fixed: initially wrong predictions made correct (out of 343/339/353 errors). Broken: initially correct predictions made wrong (out of 1346/1350/1336). ΔAcc: change in accuracy within the indicated group, mean $\pm$ sample standard deviation across seeds; "all" uses all 1689 examples. Base accuracy is $.796 \pm .004$.

			ΔAcc		
Rule	Fixed	Broken	all	$m \leq 1$	$m > 1$
Variance	111 / 111 / 78	212 / 223 / 159	$-.058 \pm .009$	$-.050 \pm .014$	$-.061 \pm .015$
Gradient ($\hat{y}$)	120 / 109 / 98	264 / 255 / 191	$-.076 \pm .018$	$-.065 \pm .018$	$-.080 \pm .021$
Gradient (gold)	123 / 122 / 104	264 / 255 / 191	$-.071 \pm .017$	$-.057 \pm .016$	$-.077 \pm .022$
PL	109 / 84 / 89	253 / 227 / 207	$-.080 \pm .009$	$-.078 \pm .017$	$-.081 \pm .009$
Weight-only	119 / 123 / 81	267 / 210 / 214	$-.073 \pm .019$	$-.049 \pm .013$	$-.082 \pm .021$
Random	94 / 102 / 75	224 / 222 / 154	$-.065 \pm .016$	$-.071 \pm .013$	$-.063 \pm .017$

we have not tested whether any difference between rules is real. As expected, the two gradient scores behave identically on correct predictions, where the predicted and true labels agree.

Replacement matters most on flagged instances, in both directions: it fixes more of their errors, but it also breaks more of their correct predictions (Table 11). Flagged instances hold just over half of the errors, yet most of the fixes happen there. This is what a sensitivity diagnostic should show. It does not make the ranking a way to improve accuracy.

We do not know why replacement hurts. One possibility is that the task head was trained on averaged concept predictions but receives exact 0 or 1 values after replacement; how a bottleneck is trained affects how it responds to interventions (Koh et al., 2020). Most replacements also land outside the ellipsoid, especially for flagged instances (Appendix M.2). For certified instances, any change in prediction must come from such an out-of-set replacement, as the certificate implies.

Before any replacement, the rules often pick the same concept (Table 10). PL agrees most often with the gradient score and less often with variance, and agreement is partly automatic when only one aspect is eligible. So the rules frequently choose alike, and all of them lose accuracy. Whether two rankings agree and whether their replacements help are separate questions.

Training and ensemble construction. A key point from training is that robust training changes the uncertainty set more reliably than it changes accuracy (Table 7 in Appendix G). On CEBaB, Mode C roughly doubles the width of the uncertainty set compared with Mode A, while accuracy stays about the same. On CUB, the standard Mode C setup is less accurate than Mode A. The two-stage setup is the most accurate of our models, but it changes initialization, loss weights, and training length at once, so we cannot credit the gain to robust training. A standard deep-ensemble CBM is more accurate still. On GoEmotions, Mode C brings no accuracy gain. Across datasets, the link between margin and width also changes sign; the margin theorem holds for each input and predicts no particular sign for this correlation.

Training also changes how the uncertainty is spread across concepts. On CUB, the two-stage model has a much smaller uncertainty set than the baseline, yet its per-concept scores are more uneven: a few concepts carry most of the sensitivity (Table 7). A smaller set can therefore concentrate its sensitivity on fewer concepts. This holds whether the scores use the true label or the predicted one.

The variance floor gives a second check (Table 8, Appendix K). With the log-barrier on, almost no concept variances hit the floor. With it off, a small share do on CUB, but most stay above it even

though the sets are smaller. Across several zero-barrier CUB runs, the average width barely moves when the robust weight or warmup changes. A stable average does not mean the same concepts stay uncertain. These runs differ in several settings at once, so they describe what happened; they do not isolate the effect of any one setting.

Two CEBaB ablations examine the sources of ensemble diversity (Table 12). E1a uses independent initializations with common dropout .175; E1b additionally initializes all heads identically, while stochastic training can still make them diverge. Both use five seeds, compared with three for the original ensemble. Accuracy, width, and margin MRA are similar across configurations; the latter is .857, .855, and .855 on the common coverage grid. These aggregate results concern one dataset and do not establish equivalent rankings or routing decisions.

Additional analyses. The ablations expose that no single training choice helps everywhere (Appendix N). In single-seed runs, a larger robust weight gives the best accuracy on both CEBaB and CUB, and the stationarity penalty helps a little (Table 13). Stronger dropout, however, helps slightly on CEBaB but clearly hurts on CUB, so adding more randomness during training is not always good.

Modeling correlations between concepts did not help either (Table 14). The additional diagonal and low-rank runs have similar widths to each other and are both less accurate than the original diagonal baseline, and they differ from it in other settings too. With only five heads, an estimated covariance is at most rank four, so how it is estimated and regularized matters as much as its shape.

The best encoder also depends on the task (Table 15). RoBERTa is best on CEBaB, while DistilBERT is better on GoEmotions, even though both have nearly the same width there. Models with the same amount of disagreement can therefore differ in accuracy. For this reason, the audit should be run on each trained model; neither width nor a training recipe is a model-independent measure of reliability.

5 Related Work

Concept uncertainty and intervention. CBMs expose concepts that can be corrected before the task prediction is recomputed (Koh et al., 2020). Probabilistic CBMs represent uncertainty through concept embeddings (Kim et al., 2023), while stochastic CBMs model concept dependence and propagate an intervention through correlated concepts (Vandenhirtz et al., 2024). Post-hoc stochastic CBMs add a learned covariance model to an existing CBM (Hoffmann et al., 2026). Our focus is the relationship between a specified concept-space region and the decision boundaries of a linear task head. We use ensemble dispersion to define that region and derive a deterministic certificate conditional on it; we do not interpret the region as a distribution with a calibrated coverage probability.

Intervention selection also depends on the objective being optimized. Shin et al. (2023) study uncertainty, gradient-contribution and expected-prediction-change criteria, together with intervention costs and granularity. The confidence-region strategy of Vandenhirtz et al. (2024) uses a distributional model to choose intervention values. Our attribution instead measures each axis's contribution to a fixed squared sensitivity penalty. Its Shapley interpretation concerns this surrogate, whereas the correction experiments measure the effect of replacing a concept by its annotation. Distinguishing these objectives is essential when a sensitivity ranking fails to improve task accuracy.

Robust optimization and credal predictions. Distributionally robust optimization minimizes worst-case expected loss over a set of data distributions (Hu & Hong, 2012; Blanchet et al., 2024). CreDRO constructs credal predictions from ensembles trained under different robust reweightings (Wang et al., 2026). Our uncertainty set instead contains concept vectors for an individual input. The connection is the use of set-valued uncertainty and robust objectives. The ellipsoidal support calculation gives the exact maximum of a linearized loss, while the prediction-invariance certificate is exact for affine logits on the unboxed ellipsoid. Neither result replaces nonlinear adversarial optimization (Madry et al., 2019; Sinha et al., 2020) or supplies a population-level robustness guarantee.

Selective prediction and deferral. Selective classification studies the trade-off between retained coverage and accuracy (Geifman & El-Yaniv, 2017). Learning to defer additionally accounts for consultation with an expert (Madras et al., 2018; Mozannar & Sontag, 2021), while active learning studies informative queries under a labeling budget (Settles, 2009). We evaluate the margin as a

selective-prediction score and report the partition induced by its geometric threshold. This threshold does not encode an expert's accuracy, a query cost, or a target error rate. A complete deferral policy would need these quantities in addition to the conditional sensitivity information provided here.

6 LIMITATIONS

The certificate depends on the chosen ensemble-disagreement region and an affine task head. Ensemble members can share systematic errors, and a stable decision can remain wrong because concepts are inaccurate, missing, or insufficient for the task. Scaling the covariance changes the threshold partition; the unit ellipsoid is a modeling choice, not a calibrated confidence region. The support calculation for a fixed local gradient extends to differentiable nonlinear heads, but an exact nonlinear prediction certificate requires additional control that we do not establish.

The attribution assumes diagonal geometry and a fixed gradient. An observed correction changes the concept mean and may change the predicted class and gradient. The resulting score does not optimize post-correction accuracy, unequal query costs, or sequential interventions. Our intervention evaluation is restricted to single-concept replacement on CEBaB and shows a net accuracy loss under every rule. It does not identify the cause of that loss. The remaining datasets support accuracy and routing analyses, and several baseline and training comparisons have uneven seed counts or change multiple factors. These restrictions limit conclusions about the benefit of joint training and the generality of the intervention findings. No human-subject evaluation has been conducted.

7 CONCLUSION

We studied how disagreement among concept predictors affects a CBM's decision. An ensemble-defined ellipsoid yields a closed-form certificate of prediction invariance for a linear task head, together with a decomposition of local loss sensitivity across concepts for diagonal geometry. These diagnostics separate the amount of concept disagreement from its relevance to a particular decision. The experiments support the margin as a competitive selective-prediction score, while the CEBaB correction results show that sensitivity and intervention benefit can diverge. An audit can therefore establish stability within a specified region and identify contributions to local sensitivity without implying that a prediction is correct or that replacing a concept will improve it.

AI USE STATEMENT

We used generative AI tools to gather feedback on writing parts and editing of 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

Appendices C, D, and E provide assumptions and complete proofs. Appendix A records the available model configurations. Section 4 and Appendices H and M.2 define the evaluation metrics and replacement protocol. The empirical appendices distinguish documented settings from details that remain to be verified.

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.

Jose Blanchet, Jiajin Li, Sirui Lin, and Xuhui Zhang. Distributionally robust optimization and robust statistics, 2024. URL https://arxiv.org/abs/2401.14655.

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.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Wiktor Jan Hoffmann, Sonia Laguna, Moritz Vandenhirtz, Emanuele Palumbo, and Julia E. Vogt. Post-hoc stochastic concept bottleneck models, 2026. URL https://arxiv.org/abs/2510.08219.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2012.

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.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018. URL https://arxiv.org/abs/1711.06664.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert, 2021. URL https://arxiv.org/abs/2006.01862.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, Princeton, 1953.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31504–31520. PMLR, 2023. URL https://proceedings.mlr.press/v202/shin23a.html.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training, 2020. URL https://arxiv.org/abs/1710.10571.

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia E. Vogt. Stochastic concept bottleneck models, 2024. URL https://arxiv.org/abs/2406.19272.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Dataset*, juill 2011.

Kaizheng Wang, Ghifari Adam Faza, Fabio Cuzzolin, Siu Lun Chau, David Moens, and Hans Hallez. Learning credal ensembles via distributionally robust optimization, 2026. URL https://arxiv.org/abs/2602.08470.

A Notation and Reproducibility

Table 5: Core notation. Encoded concept annotations belong to $[0, 1]$; raw categorical labels require an explicit encoding.

Symbol	Meaning
$\mathbf{x}, y$	Input and task label
K, H, C	Number of concepts, ensemble heads, and task classes
B	Number of concepts selected for inspection
$\mathbf{p}^{(h)}, \boldsymbol{\mu}$	Head predictions and their mean
c_k	Encoded annotation used for concept replacement
$\widehat{\sigma}_k^2$	Raw across-head sample variance
$\boldsymbol{\Sigma}_{\mathrm{epi}}, \sigma_k^2$	Shape matrix and effective diagonal variances
$\widetilde{\mathcal{C}}, \mathcal{C}$	Unboxed ellipsoid and probability-box intersection
$\epsilon, \bar{\epsilon}$	Trace-based width and its test-set mean
$\mathbf{W}, \mathbf{b}$	Linear task-head parameters
$\hat{y}, \mathbf{q}$	Nominal predicted class and softmax probabilities
d_c, s_c, m_c, m	Pairwise gap, directional scale, class margin, minimum margin
$\mathbf{g}, a_k$	Local loss gradient and squared-penalty contribution
a_k^{PL}	Predicted-logit comparator
$\lambda_c, \lambda_{\mathrm{dro}}, \beta$	Concept, robust-surrogate, and barrier weights

Recorded configuration. The recorded ensemble has $H = 5$ concept MLPs with hidden width 256. The default training dropout rates span $[0.05, 0.30]$, while the matched-dropout ablations use 0.175. Per-concept standard deviations are clamped to $[10^{-4}, 2]$, corresponding to a variance floor of 10^{-8}. The recorded CEBaB runs use batch size 16 and learning rate 10^{-3}; E1a and E1b train for 50 epochs and the two-stage runs for 70 epochs. Recorded CUB runs use batch size 64 and learning rate 10^{-4}, with the two-stage schedule detailed in Appendix I. The CUB baseline description specifies gradient clipping at 1 and a five-epoch ramp of the robust term; the different warmup entry in Table 8 requires the clarification recorded in the empirical appendix.

Encoders are frozen and features are cached. The 35 runs with timing summaries report 20 s to 22 min per run and 2.2 h in total. The manuscript reports an inventory of 88 runs; the manifest and its anonymous supplementary location remain to be verified. Hardware and the historical training commit were not recorded, nor were frozen example IDs or evaluation-time dropout for the original evaluations. The CEBaB re-evaluation uses IDs recovered from the raw records and disables dropout. These provenance gaps are recorded with the remaining empirical checks.

B Alternative Ambiguity-Set Geometries

The uncertainty region determines the support function used by both diagnostics. We give the unboxed expressions below and distinguish them from their probability-box intersections.

Full-covariance ellipsoids. For any positive semidefinite $\boldsymbol{\Sigma}$, the centered support is $\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}}$. Thus Theorems 1 and 2 also apply to full and singular matrices under the square-root definition in Equation equation 1. The squared penalty now includes cross terms $2\Sigma_{ij} g_i g_j$. Extending the attribution requires a game that specifies how inspecting a correlated concept changes the joint region.

Low-rank covariance. A matrix $\boldsymbol{\Sigma} = \mathbf{D} + \mathbf{V}\mathbf{V}^\top$, with diagonal $\mathbf{D} \succeq 0$ and $\mathbf{V} \in \mathbb{R}^{K \times r}$, gives

$$\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} = \mathbf{a}^\top \mathbf{D} \mathbf{a} + \|\mathbf{V}^\top \mathbf{a}\|_2^2.$$

This allows efficient evaluation of both diagnostics. A rank-r truncation of a sample covariance is the special case $\mathbf{D} = 0$. Neither construction is generally axis-aligned. Since a sample covariance from H heads has rank at most $H - 1$, a full-matrix implementation must handle singularity or report its regularization. A positive diagonal regularizer changes the region.

Hyperrectangles. For $\mathcal{B} = \prod_k [\mu_k - \delta_k, \mu_k + \delta_k]$, the linearized penalty is $\sum_k \delta_k |g_k|$. A maximizer displaces every coordinate with nonzero g_k to the corresponding endpoint, rather than concentrating on a single coordinate. The pairwise tie hyperplane for class c is reachable if and only if $d_c \leq \sum_k \delta_k |w_{c,k}|$. Fixing the local gradient and removing individual intervals also gives an additive game, with contribution $\delta_k |g_k|$. Thus, tractable margins and additive attributions are not unique to ellipsoids. For box-clipped intervals, let $l_k = \max(-\delta_k, -\mu_k)$ and $u_k = \min(\delta_k, 1 - \mu_k)$. The exact clipped support in direction $\mathbf{a}$ is $\sum_k \max(a_k l_k, a_k u_k)$.

C Proof of Theorem 1

For brevity we write $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{epi}}$. Every displacement in $\widetilde{\mathcal{C}} - \boldsymbol{\mu}$ has the form $\boldsymbol{\Sigma}^{1/2}\mathbf{u}$ with $\|\mathbf{u}\|_2 \leq 1$. Therefore

$$\max_{\mathbf{p} \in \widetilde{\mathcal{C}}} \mathbf{g}^\top (\mathbf{p} - \boldsymbol{\mu}) = \max_{\|\mathbf{u}\|_2 \leq 1} (\boldsymbol{\Sigma}^{1/2}\mathbf{g})^\top \mathbf{u} = \|\boldsymbol{\Sigma}^{1/2}\mathbf{g}\|_2 = s.$$

If $s > 0$, equality in Cauchy–Schwarz is attained at $\mathbf{u} = \boldsymbol{\Sigma}^{1/2}\mathbf{g}/s$, which yields $\mathbf{p}^{\text{lin}} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{g}/s$. If $s = 0$, all feasible displacements have zero inner product with $\mathbf{g}$. Restricting to $\mathcal{C} \subseteq \widetilde{\mathcal{C}}$ can only reduce the maximum; it leaves the value unchanged whenever $\mathbf{p}^{\text{lin}}$ is box-feasible. The argument requires neither an invertible covariance nor an affine task head: it holds for the local gradient of any differentiable loss.

Singular covariance. The square-root definition is equivalently

$$\widetilde{\mathcal{C}} = \{\boldsymbol{\mu} + \boldsymbol{\delta} : \boldsymbol{\delta} \in \text{range}(\boldsymbol{\Sigma}),\ \boldsymbol{\delta}^\top \boldsymbol{\Sigma}^\dagger \boldsymbol{\delta} \leq 1\},$$

where $\boldsymbol{\Sigma}^\dagger$ is the Moore–Penrose pseudoinverse. The range condition is essential: a pseudoinverse constraint alone would leave null-space displacements unconstrained. For a diagonal matrix, any zero-variance coordinate is fixed at its mean.

Taylor remainder. Let A be either $\widetilde{\mathcal{C}}$ or $\mathcal{C}$, and define $h_A(\mathbf{a}) = \max_{\mathbf{p} \in A} \mathbf{a}^\top (\mathbf{p} - \boldsymbol{\mu})$. Suppose ℓ is twice differentiable and $\|\nabla^2 \ell(\mathbf{p})\|_{\text{op}} \leq L$ on $\widetilde{\mathcal{C}}$. Taylor's theorem and $\|\mathbf{p} - \boldsymbol{\mu}\|_2^2 \leq \lambda_{\max}(\boldsymbol{\Sigma})$ give

$$\left| \max_{\mathbf{p} \in A} \ell(\mathbf{p}) - \big(\ell(\boldsymbol{\mu}) + h_A(\mathbf{g})\big) \right| \leq \frac{L}{2} \lambda_{\max}(\boldsymbol{\Sigma}).$$

For cross-entropy with a linear task head, ℓ is convex, so its first-order approximation is a lower bound. Consequently,

$$0 \leq \max_{\mathbf{p} \in A} \ell(\mathbf{p}) - \ell(\boldsymbol{\mu}) - h_A(\mathbf{g}) \leq \frac{L}{2} \lambda_{\max}(\boldsymbol{\Sigma}).$$

The closed-form value is usually not the true worst-case cross-entropy. It does bound the tangent-line loss once the box is applied, but it does not bound the real loss inside the box unless we also account for the approximation error.

Exact clipped support for a diagonal covariance. Let $I = \{k : \sigma_k^2 > 0\}$ and fix all other displacements to zero. The clipped support in direction $\mathbf{a}$ solves

$$h_{\mathcal{C}}(\mathbf{a}) = \max_{\boldsymbol{\delta}} \sum_{k \in I} a_k \delta_k \quad \text{subject to} \quad \sum_{k \in I} \frac{\delta_k^2}{\sigma_k^2} \leq 1, \qquad -\mu_k \leq \delta_k \leq 1 - \mu_k.$$

First choose the box endpoint in the direction of a_k, taking $\delta_k = 0$ if $a_k = 0$. If this vector satisfies the ellipsoid constraint, it is optimal. Otherwise there is a scalar $\tau > 0$ such that

$$\delta_k(\tau) = \text{clip}\left(\frac{\sigma_k^2 a_k}{\tau}, -\mu_k, 1 - \mu_k \right), \qquad \sum_{k \in I} \frac{\delta_k(\tau)^2}{\sigma_k^2} = 1. \tag{9}$$

The sum is continuous and nonincreasing in τ, so a one-dimensional search finds such a value. To verify optimality, $\boldsymbol{\delta}(\tau)$ maximizes $\mathbf{a}^\top \boldsymbol{\delta} - (\tau/2) \sum_k \delta_k^2 / \sigma_k^2$ over the box. For any feasible $\boldsymbol{\delta}'$, this yields

$$\mathbf{a}^\top \boldsymbol{\delta}' \leq \mathbf{a}^\top \boldsymbol{\delta}(\tau) + \frac{\tau}{2} \left(\sum_k \frac{(\delta_k')^2}{\sigma_k^2} - 1 \right) \leq \mathbf{a}^\top \boldsymbol{\delta}(\tau).$$

Coordinatewise clipping of the unboxed solution alone need not use the available ellipsoidal budget optimally; Equation equation 9 reallocates it through τ. For a full covariance, the corresponding problem is a linear objective with a Euclidean-ball constraint and affine box constraints in $\mathbf{u}$, and can be solved as a second-order cone program.

D Proof of Theorem 2

For $c \neq \hat{y}$, write $d_c = z_{\hat{y}}(\boldsymbol{\mu}) - z_c(\boldsymbol{\mu}) \geq 0$ and $s_c = \sqrt{\mathbf{w}_c^\top \boldsymbol{\Sigma}_{\text{epi}} \mathbf{w}_c}$. The support calculation in Appendix C, applied to $-\mathbf{w}_c$, gives

$$\min_{\mathbf{p} \in \widetilde{\mathcal{C}}} \{z_{\hat{y}}(\mathbf{p}) - z_c(\mathbf{p})\} = d_c - s_c.$$

The gap equals $d_c \geq 0$ at $\boldsymbol{\mu}$. By convexity of the set and continuity of the gap, a point with zero gap exists if and only if $d_c - s_c \leq 0$. The zero-denominator conventions in Definition 1 make this equivalent to $m_c \leq 1$, including a tie that is constant throughout a singular ellipsoid. Requiring $d_c - s_c > 0$ for every competitor is exactly the condition that $\hat{y}$ remains the unique highest-logit class throughout $\widetilde{\mathcal{C}}$. This is equivalent to $m > 1$ and implies the same property on the subset $\mathcal{C}$.

Exact boxed condition. Using the centered support $h_{\mathcal{C}}$ defined above,

$$\min_{\mathbf{p} \in \mathcal{C}} \{z_{\hat{y}}(\mathbf{p}) - z_c(\mathbf{p})\} = d_c - h_{\mathcal{C}}(-\mathbf{w}_c).$$

Thus, the boxed set reaches the pairwise tie hyperplane if and only if $d_c \leq h_{\mathcal{C}}(-\mathbf{w}_c)$; unique prediction invariance holds if and only if all these inequalities are strict in the opposite direction. Since $h_{\mathcal{C}}(-\mathbf{w}_c) \leq s_c$, the closed-form test $m > 1$ is sufficient, but can be conservative.

When $s_c > 0$, the tie point nearest to $\boldsymbol{\mu}$ in the covariance-induced Mahalanobis metric is

$$\mathbf{p}_c^{\text{tie}} = \boldsymbol{\mu} - \frac{d_c}{s_c^2} \boldsymbol{\Sigma}_{\text{epi}} \mathbf{w}_c. \tag{10}$$

More precisely, its displacement minimizes $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_{\text{epi}}^\dagger \boldsymbol{\delta}$ over $\boldsymbol{\delta} \in \text{range}(\boldsymbol{\Sigma}_{\text{epi}})$ subject to $\mathbf{w}_c^\top \boldsymbol{\delta} = -d_c$. The minimum is m_c^2; this is not generally the Euclidean nearest point. If $m_c \leq 1$ and $\mathbf{p}_c^{\text{tie}}$ is box-feasible, it witnesses a tie in $\mathcal{C}$. This feasibility check is sufficient, since another tie point may be feasible when $\mathbf{p}_c^{\text{tie}}$ is not. A third class can also dominate a pairwise tie; the aggregate strict-gap condition accounts for all competitors.

Why the boxed converse fails. Consider one concept, $\mu = 0.1$, $\sigma = 0.3$, and the pairwise gap $z_{\hat{y}}(p) - z_c(p) = p + 0.1$. Here $m = 0.2/0.3 < 1$, and the unboxed interval $[-0.2, 0.4]$ reaches the tie at $p = -0.1$. Its clipped version $[0, 0.4]$ has strictly positive gap everywhere. This example concerns the specified geometry, rather than an experimental observation.

Scope of the certificate. At a fixed input, multiplying covariance by t^2, for $t > 0$, divides nondegenerate margins by t. Across inputs, nominal gaps and covariance orientations also vary, so the theorem imposes no sign on the correlation between width and margin. Its invariance statement applies only within $\mathcal{C}$; an oracle correction outside the set can change a prediction with $m > 1$.

Empirical box feasibility. Table 6 distinguishes the linearized-loss maximizer from the nearest tie point for the competitor attaining m. Feasibility of the former makes the unboxed linearized value exact on the boxed set. Feasibility of the latter witnesses a boxed tie only when $m \leq 1$; when $m > 1$, it is outside the ellipsoid. Among flagged instances, the tie point is feasible for 494 of 495 on CEBaB and for every flagged instance in both CUB runs. Across 72 audited diagonal runs, an exact boxed solve certified 4 additional instances and never contradicted $m > 1$. These checks use feasibility tolerance 10^{-9} and do not establish a general boxed converse.

E Shapley Values of the Fixed Local Game

Let $a_k = \sigma_k^2 g_k^2$ and $A = \sum_k a_k$, with the local gradient and covariance fixed. Define $v(S) = \sum_{k \in S} a_k$. Adding player k to any coalition that excludes it changes the value by a_k. Averaging this

Table 6: Box feasibility on seed-0 runs, with explicit denominators. "Gold" and "pred." refer to the label used to form $\mathbf{g}$.

Run	$\mathbf{p}^{\mathrm{lin}}$ gold / all	$\mathbf{p}^{\mathrm{lin}}$ pred. / all	Tie / all	Tie / $m \leq 1$
CEBaB, Mode C	1650/1689	1669/1689	1148/1689	494/495
CUB, Mode C baseline	5794/5794	5794/5794	3492/5794	1797/1797
CUB, Mode C two-stage	5760/5794	5759/5794	1125/5794	339/339

marginal contribution over all player orderings gives $\phi_k = a_k$, proving Proposition 1. Equivalently, the residual game $A - v(S)$ has Shapley contributions $-a_k$, with the same ranking of magnitudes.

Removing coordinate k changes the unsquared first-order penalty from $\sqrt{A}$ to $\sqrt{A - a_k}$. The reduction is increasing in a_k, so the same coordinate maximizes a single-step reduction in either quantity. However, a_k is the Shapley value of the *squared*-penalty game. The marginal reduction in the unsquared game depends on the other coordinates already removed; its Shapley values are not generally a_k.

The largest B values also maximize $\sum_{k \in S} a_k$ among sets of size B in this fixed game. Sequential corrections change the mean, gradient, and potentially the class. Their utility additionally depends on the correction $c_k - \mu_k$, the nominal-loss change, oracle error, and query costs, none of which this game models.

F Comparison of Axis-Selection Rules

Variance ranking measures disagreement, whereas the loss-gradient score attributes a fixed squared loss penalty. The PL comparator, $\sigma_k^2 W_{\hat{y},k}^2$, instead attributes the squared ellipsoidal sensitivity of $z_{\hat{y}}(\mathbf{p})$. It has an additive Shapley interpretation for that logit game, which is distinct from the cross-entropy game in Proposition 1.

For cross-entropy,

$$\mathbf{g} = \mathbf{W}^\top(\mathbf{q} - \mathbf{e}_y) = \sum_c q_c \mathbf{W}_c - \mathbf{W}_y.$$

If $y = \hat{y}$ and $q_{\hat{y}}$ approaches one, $\mathbf{g}$ approaches zero and its direction depends on the competing class weights. If $y \neq \hat{y}$, its limit is $\mathbf{W}_{\hat{y}} - \mathbf{W}_y$. Neither case generally supports replacing $\mathbf{g}$ by $\mathbf{W}_{\hat{y}}$. Our deployable gradient rule fixes $y = \hat{y}$ at the nominal prediction. Table 4 compares this rule with the retrospective gold-label gradient, PL, variance, weight-only, and random rankings.

Predicted-logit weighting also depends on the representation of the logits. Replacing every row $\mathbf{W}_c$ by $\mathbf{W}_c + \mathbf{a}^\top$ adds the same input-dependent value to every logit, leaving softmax probabilities, predictions, cross-entropy gradients, and pairwise margins unchanged. It can nevertheless change $W_{\hat{y},k}^2$ and hence the PL ranking. PL should therefore be treated as a diagnostic of the fitted logit parameterization, not an invariant measure of decision influence.

Variance ranking agrees with product ranking when squared sensitivities are constant across concepts; sensitivity-only ranking agrees when variances are constant. These conditions are sufficient, not necessary. The fixed-game result establishes neither rule's superiority in realized correction accuracy. Our weight-only comparator, $\max_c |W_{c,k}|$, gives a fixed concept ordering before the eligibility mask is applied; random ranking provides a further empirical control.

G Joint Training: Derivations and Diagnostics

We distinguish the numerical surrogate from its differentiation convention and then derive a variance equalization condition under explicit simplifying assumptions.

Table 7: Training-mode diagnostics. n is the number of seeds; $\pm$ denotes standard deviation. Width and margin–width correlation describe the learned geometry. CV is the mean per-instance coefficient of variation (ddof $= 0$) of $\sigma_k^2 g_k^2$ across concepts, using the gold-label gradient. It measures unequal local scores, rather than convergence of the neural objective. Recomputation gives 1.142 on baseline CUB (1.131 with the predicted label). Mode A/B values are unavailable because the saved gradients were zero.

DATASET	MODE	N	ACC	$\bar{\epsilon}$	$\rho(\epsilon, m)$	CV
CEBaB (K=4)	A	3	0.789 ± 0.006	0.144 ± 0.030	-0.731 ± 0.073	—
	B	3	0.788 ± 0.005	0.166 ± 0.016	-0.772 ± 0.065	—
	C, baseline	3	0.796 ± 0.004	0.276 ± 0.009	-0.493 ± 0.012	1.014 ± 0.185
GoEmotions (K=28)	A	1	0.446	0.318	-0.299	—
	B	1	0.446	0.301	-0.295	—
	C, baseline	1	0.445	1.189	$+0.320^*$	1.299
CUB (K=112)	A	3	0.857 ± 0.001	0.838 ± 0.006	-0.031 ± 0.015	—
	B	3	0.809 ± 0.002	3.319 ± 0.004	-0.193 ± 0.011	—
	C, baseline	1	0.817	3.252	-0.177	1.142
	C, two-stage	3	0.878 ± 0.002	0.899 ± 0.008	-0.399 ± 0.026	3.027 ± 0.019

* Positive observed correlation; see §4. Two-stage Mode C: Mode A pretraining, then Mode C fine-tuning with β=0 and positive-weighted concept BCE (Appendix I).

G.1 FIRST-ORDER SURROGATE AND ITS VALIDATION

For $s = \sqrt{\mathbf{g}^\top \mathbf{\Sigma}_{\text{epi}} \mathbf{g}} > 0$, Theorem 1 gives

$$\mathbf{p}^{\text{lin}} = \boldsymbol{\mu} + \frac{\mathbf{\Sigma}_{\text{epi}} \mathbf{g}}{s}, \qquad \mathcal{L}_{\text{lin}} = \ell(\boldsymbol{\mu}) + s. \tag{11}$$

At $s = 0$, take $\mathbf{p}^{\text{lin}} = \boldsymbol{\mu}$. The value $\mathcal{L}_{\text{lin}}$, nonlinear cross-entropy at $\mathbf{p}^{\text{lin}}$, and the exact boxed maximum are distinct quantities. We do not measure their gaps; Appendix C provides the remainder bound. All reported Mode C runs use the first-order surrogate rather than projected-gradient ascent.

Differentiation convention. The available implementation forms the gradient at a detached center, retaining its dependence on the classifier parameters. Writing sg for stop-gradient, this amounts to

$$\widetilde{\mathbf{g}}_{y,\theta} = \nabla_{\mathbf{p}} \, \text{CE}(\mathbf{W}_\theta \mathbf{p} + \mathbf{b}_\theta, y)\big|_{\mathbf{p}=\text{sg}(\boldsymbol{\mu}_\theta)} . \tag{12}$$

Its forward value equals $\mathbf{g}_y$, but its derivative omits the dependence through $\boldsymbol{\mu}_\theta$. Concept heads can therefore receive gradients through covariance in the sensitivity term, while the gradient-through-center pathway is absent. Equation equation 8 records the scalar loss value; it does not specify full differentiation of $\mathbf{g}_y(\boldsymbol{\mu}_\theta)$.

In our implementation, the cross-entropy term inside $\mathcal{L}_{\text{FO}}$ is evaluated at $\boldsymbol{\mu}_\theta$ itself; only the copy of $\boldsymbol{\mu}$ used to form $\mathbf{g}$ is detached. The concept heads therefore receive the nominal cross-entropy with total weight $1 + \lambda_{\text{dro}}$, scaled by the ramp factor early in training.

Effective variances and numerical settings. The recorded diagonal implementation uses $\sigma_k^2 = \text{clip}(\sqrt{\widehat{\sigma}_k^2}, 10^{-4}, 2)^2$, where $\widehat{\sigma}_k^2$ is the raw sample variance. These effective variances define the regularized shape matrix. The quadratic form in the surrogate is additionally clamped below at 10^{-12}. The unfloored barrier $-\sum_k \log \sigma_k^2$ diverges as a variance tends to zero; with the recorded hard floor, the penalty is bounded and supplies no derivative through the clamp below the floor. Thus its anti-collapse effect above the floor does not guarantee recovery of a collapsed axis.

During the initial ramp, λ_{dro} and β are multiplied by a factor in $[0, 1]$; Mode A uses neither term. In Mode B, the robust term updates only the task classifier. The model has no aleatoric head. The stationarity variant adds $\text{mean}_k(a_k / \max(\bar{a}, 10^{-12}) - 1)^2$, with $a_k = \sigma_k^2 \text{sg}(g_k)^2$ and $\bar{a}$ the stop-gradient mean of the a_k. These statements describe the available source, without establishing the identity of the historical training commit.

Mode B (fixed radius). Mode B replaces the covariance-shaped set with a Euclidean ball of fixed radius around the concept mean,

$$\mathcal{C}_{\text{B}}(\mathbf{x}) = \{\mathbf{p} \in [0, 1]^K : \, \|\mathbf{p} - \boldsymbol{\mu}_\theta(\mathbf{x})\|_2 \leq r\},$$

with no covariance shaping and no normalization by K. The radius is $r = 0.3$ on CEBaB and $r = 3.25$ on CUB. The inner maximization is approximated by ten gradient-ascent steps of size 0.01 from $\mathbf{p}^{(0)} = \boldsymbol{\mu}_\theta$, without random restarts or gradient normalization. Each step projects onto the ball and then clips to $[0,1]^K$; because $\boldsymbol{\mu}_\theta \in [0,1]^K$, clipping keeps the iterate in the ball, so every iterate is feasible, although this is not the exact projection onto $\mathcal{C}_{\mathrm{B}}$. During the search, the center and classifier parameters are held fixed. The robust term is $\ell_y\big(\mathbf{W}\,\mathrm{sg}(\mathbf{p}^{(10)}) + \mathbf{b}\big)$, with gradients flowing to $\mathbf{W}$ and $\mathbf{b}$ but not through $\mathbf{p}^{(10)}$, so this term updates only the classifier. The Mode B objective is

$$\ell_y(\boldsymbol{\mu}_\theta) + \lambda_c \mathcal{L}_{\mathrm{concept}} + \lambda_{\mathrm{dro}}\, \ell_y\big(\mathbf{W}\,\mathrm{sg}(\mathbf{p}^{(10)}) + \mathbf{b}\big) + \beta\,\Omega,$$

with λ_{dro} and β scaled by the same ramp as in Mode C; the other terms update the concept heads as usual. Recorded runs use $\lambda_{\mathrm{dro}} = 0.1$, with $\beta = 0$ on CEBaB and $\beta = 0.005$ on CUB.

G.2 A Conditional Variance Equalization Calculation

Consider the reduced objective

$$F(\mathbf{v}) = \lambda_{\mathrm{dro}} \sqrt{\sum_k g_k^2 v_k} - \beta \sum_k \log v_k, \qquad v_k > 0,$$

where $\boldsymbol{\mu}$, $\mathbf{W}$, and $\mathbf{g}$ are fixed, the variances v_k are independent free variables, and all other loss terms are independent of $\mathbf{v}$. Assume $\lambda_{\mathrm{dro}} > 0$, $\beta > 0$, and $g_k \neq 0$ for every k. Writing $s = \sqrt{\sum_j g_j^2 v_j}$ gives

$$\frac{\partial F}{\partial v_k} = \frac{\lambda_{\mathrm{dro}} g_k^2}{2s} - \frac{\beta}{v_k}.$$

An interior stationary point therefore satisfies

$$v_k g_k^2 = \frac{2\beta s}{\lambda_{\mathrm{dro}}} \qquad \text{for every } k. \tag{13}$$

Summing over k yields $s = 2\beta K/\lambda_{\mathrm{dro}}$ and $v_k g_k^2 = 4\beta^2 K/\lambda_{\mathrm{dro}}^2$. If some $g_k = 0$, the unconstrained reduced objective has no finite interior stationary point along that coordinate. If $\beta = 0$, the robust term is nondecreasing in each v_k and favours reducing variance; Equation equation 13 no longer gives positive equalized contributions.

G.3 Interpretation for the Learned Ensemble

The neural ensemble does not satisfy the reduced problem's assumptions. Its means and variances share parameters, concept supervision affects those parameters, and probability-valued outputs constrain attainable variances. The local gradient changes during training, and variance floors affect differentiation. Network stationarity therefore need not satisfy the free-variance condition.

The coefficient of variation of $\sigma_k^2 g_k^2$ consequently describes unequal axis contributions; it is not an optimization-convergence test. The CUB two-stage configuration uses $\beta = 0$, so the positive-barrier calculation does not apply to it. Mode A/B values are unavailable because their saved gradient arrays were zero, rather than because their true local gradients vanish.

H Metric Definitions and Reporting Conventions

Task accuracy, geometric diagnostics, ranking agreement, and replacement outcomes answer different questions. We define them separately to distinguish reliability of the original prediction from the effect of changing its concept representation.

Accuracy and replacement outcomes. Task accuracy is the fraction of test instances for which $\hat{y}(\boldsymbol{\mu}) = y$. A rule r selects its highest-scoring eligible concept $k_r(\mathbf{x})$ and replaces that coordinate by the encoded annotation:

$$\boldsymbol{\mu}^{(r)}(\mathbf{x}) = \boldsymbol{\mu}(\mathbf{x}) + \big(c_{k_r}(\mathbf{x}) - \mu_{k_r}(\mathbf{x})\big)\mathbf{e}_{k_r}.$$

When no concept is eligible, the original prediction is retained. A flip is any change in predicted class; a fix changes an incorrect prediction to the true label, and a break changes a correct prediction to an incorrect one. Accuracy change is the number of fixes minus breaks, divided by the number of instances in the reported population. Fix and break rates condition on the corresponding initial outcome. Thus, a fix rate alone does not determine net accuracy change. Table 4 reports per-seed counts and mean accuracy changes; Table 11 pools counts across seeds.

Width and margin. The width is $\epsilon(\mathbf{x}) = \sqrt{\operatorname{Tr} \boldsymbol{\Sigma}_{\mathrm{epi}}(\mathbf{x})}$, and $\bar{\epsilon}$ is its test-set mean. We report Spearman correlation $\rho(\epsilon, m)$ to describe their association across inputs. Scaling a fixed ellipsoid decreases its nondegenerate margin when nominal logits remain fixed. Across inputs, logits and covariance orientations also vary, so Theorem 2 imposes no sign on this correlation.

Per-axis coefficient of variation. For $a_k = \sigma_k^2 g_k^2$, the coefficient of variation is

$$\mathrm{CV}(\mathbf{x}) = \frac{\mathrm{std}_k\, a_k(\mathbf{x})}{\mathrm{mean}_k\, a_k(\mathbf{x})},$$

When the denominator is positive, using population standard deviation across coordinates (ddof $= 0$). The reported value is its test-set mean with the gold-label gradient. It describes unequal contributions to the local penalty. It does not test stationarity of the neural objective without the assumptions in Appendix G.2. Saved Mode A/B gradients were zero, so the corresponding diagnostic is unavailable.

Ranking agreement. For top-B concept sets $R_B^r(\mathbf{x})$, overlap is $|R_B^{r_1} \cap R_B^{r_2}|/B$. Top-one agreement is the fraction of instances on which two rules select the same concept. Per-instance rank correlation is Spearman's ρ over their eligible concept scores. Its mean excludes instances with fewer than two eligible concepts or constant scores.

I Two-Stage Configuration of Mode C

The CUB two-stage configuration trains Mode A for 70 epochs, then initializes Mode C from this checkpoint and trains for another 70. The second stage uses $\lambda_{\mathrm{dro}} = 0.03$, $\beta = 0$, and positive-class weighting of 3.88 in the concept BCE, whereas baseline Mode C uses $\lambda_{\mathrm{dro}} = 0.1$ and a positive barrier. These choices aim to preserve the pretrained concept predictions while adapting the ensemble. The resulting accuracy is 0.878 ± 0.002, compared with 0.857 ± 0.001 for Mode A and 0.817 for baseline Mode C.

This comparison combines initialization, loss-weight changes, and twice the training budget of Modes A–C. It does not isolate any individual factor. Nor does pretraining establish covariance calibration. With $\beta = 0$, the positive-barrier equalization calculation in Appendix G.2 does not apply. The available source selects the earliest epoch attaining maximum validation accuracy; the historical selection rule and the derivation of the positive-class weight were not recorded. CEBaB's separately recorded two-stage runs retain $\lambda_{\mathrm{dro}} = 0.1$ and $\beta = 0.01$ (Table 8).

J Empirical Characterisation of the Ranking Rules

The replacement experiments compare variance, predicted-label gradient, retrospective gold-label gradient, PL, weight-only, and random ranking. The gradient score evaluates the local loss sensitivity in Proposition 1; PL evaluates a different, parameterization-dependent logit sensitivity. We report their agreement to show how often their selections differ, without treating agreement as evidence of intervention quality.

The present experiment replaces one concept per instance. Recomputing a ranking after each correction would define a sequential policy: the concept mean, gradient, and predicted class could all change. The fixed-game optimality statement does not determine the accuracy of such a policy.

K Variance Collapse and Training Schedule

Two forces act on the concept variances during training. The robust penalty grows with variance, so minimizing it pushes variances down. The log barrier pushes back, keeping variances from collapsing

to the floor. Table 8 shows how this balance played out in our runs. With the barrier on, almost no variances reach the floor. With it off on CUB, a small share do, and the uncertainty sets are smaller. These runs also differ in other settings, so the table shows what happened but does not prove the barrier caused it.

Table 8: Variance-floor hits and mean width on recorded runs (seed 0 unless stated). The variance floor is 10^{-8}; a slot is an (example, concept) pair. Warmup values are retained as recorded; the CUB baseline also has a separately reported five-epoch robust ramp.

Run	λ_{dro}	β	Warmup	Floor slots	$\bar{\epsilon}$
CEBaB baseline (70 ep.)	0.1	0.01	0	0/6756	0.286
CEBaB two-stage, seeds 0/1/2	0.1	0.01	0	7 / 0 / 0 of 6756	0.160 / 0.235 / 0.186
CUB baseline	0.1	0.005	0	0/648928	3.252
CUB repair	0.03	0.001	10	0/648928	1.440
CUB repair, no pos. weight	0.03	0	10	5307/648928	0.831
CUB repair	0.03	0	10	14666/648928	0.795
CUB repair	0.03	0	20	14746/648928	0.794
CUB repair	0.01	0	10	14900/648928	0.795
CUB repair	0.05	0	10	14441/648928	0.795
CUB two-stage, seeds 0/1/2	0.03	0	0	11047 / 10109 / 9910	0.890 / 0.897 / 0.910

Training uses two schedule settings. During the first W warmup epochs, the robust and barrier terms are switched off. Over the next R epochs, their coefficients rise linearly to their full values: in epoch j of the ramp they are scaled by j/R. The task and concept losses are active throughout. The CUB baseline uses $W = 0$ and $R = 5$, so its robust weight grows from 0.02 to 0.1 and its barrier weight from 0.001 to 0.005 over the first five epochs.

L Additional Baseline and Ranking Details

Table 9 reports accuracy, width, and margin–width correlation for the available baseline implementations. MC dropout uses 50 passes and deep ensembles use five models; all baseline entries have one seed. The available records do not specify the P-CBM residual branch, concept encoding, covariance estimator, or tuning protocol sufficiently to verify the implementation and match training budgets. The results therefore describe these internal baselines, without establishing a reproduction of the published probabilistic CBM.

Table 9: Task accuracy and ensemble geometry for the baselines and our three training modes. Baselines use one seed; Mode A/B/C entries with $\pm$ use three. CUB Mode C is the two-stage configuration. The baselines are our own implementations; their settings and training budgets are described in Appendix L.

Dataset	Method	Acc	$\bar{\epsilon}$	$\rho(\epsilon, m)$
CEBaB (K=4)	CBM-MCD	0.694	0.093	-0.204
	CBM-DE	0.692	0.216	$+0.172$
	P-CBM-MCD	0.751	0.088	-0.504
	P-CBM-DE	0.755	0.104	-0.681
	Mode A	0.789 ± 0.006	0.144 ± 0.030	-0.731 ± 0.073
	Mode B	0.788 ± 0.005	0.166 ± 0.016	-0.772 ± 0.065
	Mode C	0.796 ± 0.004	0.276 ± 0.009	-0.493 ± 0.012
CUB (K=112)	CBM-MCD	0.888	0.794	-0.135
	CBM-DE	0.896	2.339	-0.039
	P-CBM-MCD	0.881	0.478	-0.841
	P-CBM-DE	0.884	0.407	-0.902
	Mode A	0.857 ± 0.001	0.838 ± 0.006	-0.031 ± 0.015
	Mode B	0.809 ± 0.002	3.319 ± 0.004	-0.193 ± 0.011
	Mode C, two-stage	0.878 ± 0.002	0.899 ± 0.008	-0.399 ± 0.026
GoEmotions (K=28)	CBM-MCD	0.444	0.218	-0.233
	CBM-DE	0.450	0.653	-0.548
	P-CBM-MCD	0.468	0.197	-0.477
	P-CBM-DE	0.473	0.402	-0.579
	Mode A	0.446	0.318	-0.299
	Mode B	0.446	0.301	-0.295
	Mode C	0.445	1.189	$+0.320^{*}$

* Positive observed correlation; see §4.

Ranking agreement. On the eligible CEBaB aspects, PL agrees most with gradient weighting and less with variance and weight-only ranking (Table 10). The higher top-one agreement than mean rank correlation reflects the different questions measured by these summaries. For an instance with only one eligible aspect, all rules necessarily agree on the first choice. The corresponding CUB and GoEmotions PL implementations could not be verified, so their agreement statistics are omitted. Weight-only ranking uses $\max_c |W_{c,k}|$ and is fixed across inputs for a given classifier; its selected concept can still vary with eligibility.

Table 10: Agreement with PL on eligible CEBaB aspects, Mode C seeds 0–2 (mean $\pm$ sample standard deviation). Top-one agreement uses 1583 instances with at least one eligible aspect; Spearman correlation uses 1050 with at least two eligible aspects and nonconstant scores.

Comparator	Same top-1 (of 1583)	Mean Spearman ρ (of 1050)
Variance	$.710 \pm .039$	$.224 \pm .113$
Gradient ($\hat{y}$)	$.818 \pm .008$	$.503 \pm .034$
Gradient (gold)	$.804 \pm .019$	$.461 \pm .076$
Weight-only	$.705 \pm .059$	$.173 \pm .115$

M Supplementary Operational Analysis

The operational analyses use saved Mode C checkpoints. Rankings are formed before substituting the selected concept value. The predicted-label gradient rule uses the current prediction as its reference label; the gold-label version is a retrospective comparison. Evaluation also needs encoded concept annotations and an eligibility mask, which cannot be recovered from the task label and model outputs alone.

M.1 Selective-Prediction Summary

Margin and confidence retain the largest scores; entropy and width retain the smallest. MRA is the unweighted mean accuracy at coverages $\{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1.0\}$, with floor rounding of retained counts and input-order tie breaking. Unequal intervals in this grid mean that its average is not a conventional numerical area under a risk–coverage curve. Table 2 uses three seeds for CEBaB and one checkpoint each for CUB and GoEmotions.

The margin outperforms width on all three datasets but is slightly below confidence on CEBaB and GoEmotions and matches it on CUB. At the threshold $m = 1$, Table 3 reports how much the model keeps, how accurate it is on what it keeps, how many instances go to review, and what share of errors those contain. This threshold comes from the geometry; it is not chosen to reach a target error rate.

M.2 Population-Level Oracle Correction

Table 11 pools the CEBaB counts over three seeds, distinguishing corrections among initial errors from breaks among initially correct predictions. Each seed evaluates the same test population, so these are repeated model evaluations rather than independent additional test examples.

Table 11: Pooled single-concept replacement on CEBaB (three seeds). Fix rate: fixed errors / errors in the group. Break rate: broken correct predictions / correct predictions in the group.

	$m \leq 1$		$m > 1$	
Rule	Fix	Break	Fix	Break
Variance	233/585	304/834	67/450	290/3198
Gradient ($\hat{y}$)	239/585	331/834	88/450	379/3198
Gradient (gold)	250/585	331/834	99/450	379/3198
PL	208/585	319/834	74/450	368/3198
Weight-only	246/585	316/834	77/450	375/3198
Random	205/585	306/834	66/450	294/3198

Under variance ranking, replacement rarely stays inside the ellipsoid for flagged instances, and does so more often, though still for a minority, for retained ones. By Theorem 2, a replacement inside the ellipsoid cannot change a prediction with $m > 1$. So every fix or break among retained instances comes from a replacement that leaves the set. Whether a replacement leaves the set depends on how far the annotated value is from the mean, measured in units of that concept's uncertainty; a concept with zero variance is left by any change at all. The reverse does not hold either: a reachable decision boundary does not mean the annotated value crosses it, or that crossing it gives the correct label.

We also count how often the annotated concepts fall inside the uncertainty set. Some aspects are unannotated for each example, so we count an example as inside if its annotated aspects fit in the set when every unannotated aspect is set to its mean prediction; with diagonal covariance, this is the most favourable choice. Under this rule, fewer than one in five eligible CEBaB examples fall inside the set on each seed. This describes the set; it is not a coverage level the construction aims for.

N Ablation Details

Table 12 compares selected sources of ensemble diversity on CEBaB. Common dropout removes differences in dropout rate between heads; identical initialization additionally removes differences in starting weights. Separate heads can still diverge through stochastic training. The similar aggregate metrics do not establish identical predictions, equivalent rankings, or statistical equivalence.

Table 12: CEBaB ensemble ablations, mean ± standard deviation across seeds. E1a uses independent initializations and common dropout .175; E1b also uses identical initial weights. MRA uses the same seven coverage levels as Table 2.

Metric	Original (3 seeds)	E1a (5 seeds)	E1b (5 seeds)
Accuracy (%)	$79.57 \pm .43$	$79.05 \pm .59$	$79.09 \pm .29$
$\rho(\epsilon, m)$	$-.493 \pm .012$	$-.445 \pm .130$	$-.438 \pm .106$
Margin MRA	$.8575 \pm .0058$	$.8554 \pm .0057$	$.8546 \pm .0051$
$\bar{\epsilon}$	$.276 \pm .009$	$.266 \pm .011$	$.260 \pm .006$

The recorded Mode C settings are $\lambda_{\text{dro}} = 0.1$ and $\beta = 0.01$ on CEBaB, including its two-stage runs; $\lambda_{\text{dro}} = 0.1$ and $\beta = 0.005$ for the CUB baseline; and $\lambda_{\text{dro}} = 0.03$ and $\beta = 0$ for two-stage CUB, with $\lambda_c = 1$ and $H = 5$ throughout. For the CUB baseline, the logged training loss equals exactly the task loss plus the concept loss, $\lambda_{\text{dro}}\mathcal{L}_{\text{FO}}$, and $\beta\Omega$. The last two terms are phased in linearly over the first five epochs.

Loss weight and other variants. Table 13 compares single-seed Mode C variants trained for 70 epochs. A larger robust weight ($\lambda_{\text{dro}} = 0.3$ instead of 0.1) gives the best accuracy on both CEBaB and CUB. The stationarity penalty helps slightly on both. Stronger dropout helps slightly on CEBaB but clearly hurts on CUB. With one seed per variant, these differences are not shown to be repeatable, and they do not explain why the datasets respond differently.

Table 13: Mode C variants on CEBaB and CUB (70 epochs). Every row uses one seed. The $\lambda_{\text{dro}} = 0.3$ rows increase the robust weight over the baseline value 0.1.

Dataset	Variant	Seeds	Best val acc	Test acc	$\bar{\epsilon}$
CEBaB (K=4)	Joint baseline ($\lambda_{\text{dro}} = 0.1$)	1	0.806	0.797	0.286
	Joint, dropout $= 0.6$	1	0.805	0.801	0.304
	Joint, $\lambda_{\text{dro}} = 0.3$	1	0.806	**0.808**	0.245
	Joint, stationarity penalty 0.01	1	0.807	0.800	0.261
CUB (K=112)	Joint baseline ($\lambda_{\text{dro}} = 0.1$)	1	0.791	0.817	3.252
	Joint, dropout $= 0.6$	1	0.723	0.772	3.182
	Joint, $\lambda_{\text{dro}} = 0.3$	1	0.803	**0.825**	3.110
	Joint, stationarity penalty 0.01	1	0.794	0.819	3.209

Covariance parameterization. Table 14 compares the CUB baseline with two other configurations: a second diagonal one and a low-rank one. Both are less accurate than the baseline. The baseline also

uses diagonal covariance and differs from the other two in several settings, so the table compares whole training setups; it does not measure the effect of covariance shape alone. With only five heads, an estimated covariance can have rank at most four, and any regularization used to fix this also changes the uncertainty set.

Table 14: CUB covariance runs. The baseline's recorded covariance is diagonal; it differs from the other diagonal row in settings other than covariance structure. The baseline uses one seed; the other rows use three.

VARIANT	SEEDS	TEST ACC	$\bar{\epsilon}$
Joint baseline (diagonal)	1	$\mathbf{0.817}$	3.252
Joint, diagonal covariance	3	0.756 ± 0.002	3.183 ± 0.003
Joint, low-rank covariance	3	0.760 ± 0.005	3.187 ± 0.010
Fixed-ϵ baseline	3	0.809 ± 0.002	3.319 ± 0.004
Post-hoc baseline	3	0.857 ± 0.001	0.838 ± 0.006

Encoder comparisons. The best encoder depends on the task (Table 15). RoBERTa is clearly the most accurate on CEBaB, while DistilBERT does better on GoEmotions. Apart from RoBERTa on CEBaB, each result comes from a single run, so these comparisons do not show that any encoder is generally better.

Table 15: Reported encoder comparisons. Only the CEBaB RoBERTa row uses three seeds; every other row uses one.

DATASET	ENCODER	SEEDS	TEST ACC	$\bar{\epsilon}$
CEBaB (K=4)	BERT-base	1	0.737	0.165
	DistilBERT-base	1	0.731	0.157
	RoBERTa-base	3	$\mathbf{0.796 \pm 0.004}$	0.276 ± 0.009
GoEmotions (K=28)	DistilBERT-base	1	$\mathbf{0.473}$	1.179
	RoBERTa-base	1	0.445	1.189