ENTROPY SHORTCUT IN LLM CORRECTNESS PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) achieve strong performance across textual, visual, and other multimodal tasks, yet they frequently hallucinate, producing outputs that are incorrect and unsupported. Preventing such failures is difficult, hence recent works have increasingly focused on detecting them. Two promising approaches that do not modify the model parameters are: (a) sampling the model at high temperature and measuring the entropy of the sampled responses, and (b) training a linear probe on hidden states to estimate the probability of a correct response. Entropy-based detection rests on the hypothesis that sampling entropy and correctness are inversely associated. In this work, we show that a sizable fraction of mispredictions exhibit low sampling entropy. More importantly, standard linear correctness probes fail to distinguish these low-entropy mispredictions from high-entropy correct predictions. We trace this failure to an *entropy shortcut*- the probes exploit the entropy–correctness association rather than learning a correctness signal that remains stable across entropy regimes. To address this, we propose linear detectors that predict correctness with reduced *entropy shortcut*. Experiments with four LLMs and five multimodal LLMs across eight datasets, including safety-critical domains, establish the entropy shortcut as a systematic failure mode for LLM probes and demonstrate substantially more robust correctness detection across both entropy-aligned and entropy-misaligned cases.

1 INTRODUCTION

Large language models (LLMs) Achiam et al. (2023); Comanici et al. (2025); Grattafiori et al. (2024) achieve strong performance across text and multimodal tasks, yet continue to produce incorrect or unsupported responses, commonly referred to as hallucinations. Because such failures cannot be eliminated reliably, inference-time correctness estimation is needed to support abstention or routing to stronger models or human experts Ren et al. (2023); Bedi et al. (2025). Unlike conventional deep models, for which output probabilities provide a default uncertainty signal Guo et al. (2017), LLM token probabilities are poorly aligned with response-level correctness after human-alignment training Xie et al. (2024). Hence, reliable correctness estimation remains an open challenge.

Existing correctness-estimation methods are broadly divided into **black-box** and **white-box** approaches. Black-box methods sample an LLM at a high temperature, commonly $T = 1.0$, and use the semantic diversity of its responses as a correctness signal Kuhn et al. (2023); Nikitin et al. (2024); Lin et al. (2024); Nguyen et al. (2025). They rely on an inverse association between correctness and sampling entropy: incorrect greedy responses are expected to produce diverse samples and high entropy, whereas correct responses are expected to produce consistent samples and low entropy Bhattacharya et al. (2026). White-box methods instead use internal model representations and often provide stronger correctness detection. These include hidden-state correctness probes Orgad et al. (2025), the semantic entropy probe (SEP) Han et al. (2024), and probes derived from attention matrices Binkowski et al. (2025); Sriramanan et al. (2024); Binkowski et al. (2026).

In this work, we benchmark the sampling behaviour and correctness-detection performance of four LLMs and five multimodal LLMs across eight datasets spanning diverse domains. We find that correctness and sampling entropy are not universally inversely associated: correct high-entropy and incorrect low-entropy predictions (Figure 1) occur in substantial proportions. More surprisingly, correctness probes exhibit the same failure mode despite receiving sampling entropy neither as an input nor as a training label, with performance collapsing to near-chance levels on entropy-misaligned

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.

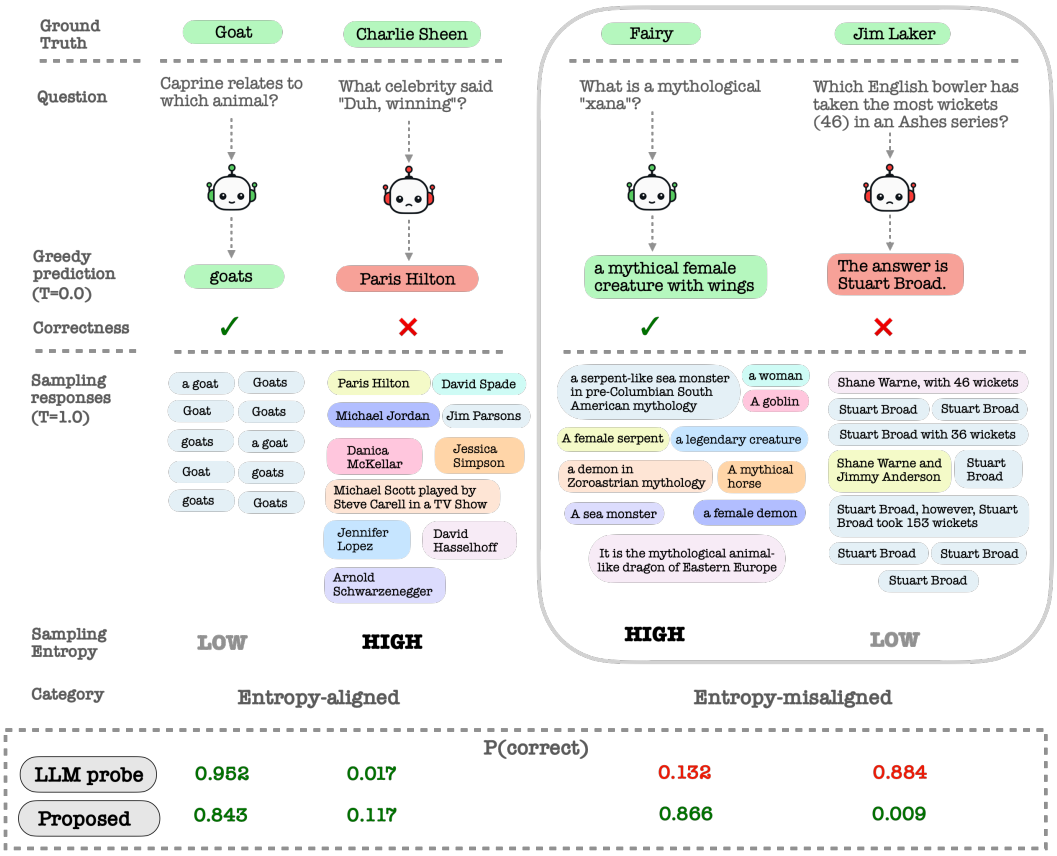


Figure 1: **Entropy shortcut in LLM correctness probes**: From left to right, the examples represent correct with low entropy (CL), incorrect with high entropy (IH), correct with high entropy (CH), and incorrect with low entropy (IL) samples. Sampling responses sharing a color are semantically same. The first two examples follow the usual entropy–correctness inverse association, whereas the rest contain conflicting cases. The standard LLM correctness probes fail on them (red colored scores), indicating entropy-shortcut. The proposed linear probes show consistent detection performance. Inferences from `Llama-3.1` (8B) Grattafiori et al. (2024) on Trivia-QA Joshi et al. (2017) are used. More examples from LLMs and multimodal LLMs across datasets are shown in Figures 5 and 6.

cases. Through entropy-band analysis and geometric investigation, we identify this behaviour as an *entropy shortcut* and propose two training paradigms to mitigate it.

Contributions:

- We show across LLMs, multimodal LLMs, and datasets from diverse domains that correctness and sampling entropy are not universally inversely associated, producing substantially correct high-entropy and incorrect low-entropy populations (**Section** 4).
- Through entropy-band evaluation and geometric analysis, we establish that linear probes trained on LLM internals exploit an entropy shortcut, creating a critical blind spot in correctness detection (**Section** 5).
- We propose two training paradigms that largely mitigate this shortcut and remain stable across entropy-aligned and entropy-misaligned cases (**Section** 6).

2 Problem Formulation

Let X denote an input to an LLM $\mathcal{M}$. Under greedy decoding ($T = 0$), the model produces $Y^0 = \mathcal{M}(X; T = 0)$. Given a ground-truth answer $Y^\star$ and an evaluator $E(\cdot, \cdot)$, we define the correctness label as $Z = E(Y^0, Y^\star) \in \{0, 1\}$, where $Z = 1$ denotes a correct response. For the same input, let $\{Y_k\}_{k=1}^K$ be responses sampled at $T = 1.0$. Following Kuhn et al. (2023), their empirical

semantic entropy is

$$\mathrm{SE}(X) = -\sum_{j=1}^{J} \widehat{p}_j(X) \log \widehat{p}_j(X), \tag{1}$$

where J is the number of semantically distinct response groups, S_j is the set of responses in group j, and $\widehat{p}_j(X) = \frac{|S_j|}{K}$. Sampling-based correctness estimation assumes that $\mathrm{SE}(X)$ is inversely associated with Z; our experiments test this assumption.

We next consider a linear correctness probe operating on the model representation $h_{\mathcal{M}}(X, Y^0) \in \mathbb{R}^d$. Its predicted probability of correctness is

$$\hat{Z} = \sigma\big(\mathbf{w}^\top h_{\mathcal{M}}(X, Y^0) + \beta\big), \tag{2}$$

where $\theta = \{\mathbf{w}, \beta\}$ denotes the probe parameters and $\sigma(\cdot)$ is the sigmoid function. Let $B \in \{L, H\}$ denote the low- or high-entropy band of a sample. An entropy shortcut arises when the probe exploits the association between B and Z and fails when that association changes Geirhos et al. (2020). Following equalized odds Hardt et al. (2016), a correctness probe without an entropy shortcut should satisfy

$$\hat{Z} \perp B \mid Z, \tag{3}$$

i.e., conditioned on correctness, its predictions should be independent of the entropy band.

3 Experimental setup

3.1 Metrics

First, consider a dataset $\mathcal{D} = \{(x_i, y_i^\star)\}_{i=1}^N$ and an LLM $\mathcal{M}$. For each sample, let $y_i^0 = \mathcal{M}(x_i; T = 0)$ and $Z_i = E(y_i^0, y_i^\star) \in \{0, 1\}$. The dataset is then partitioned into correctly predicted samples C and incorrectly predicted samples I:

$$C = \{x_i \mid (x_i, y_i^\star) \in \mathcal{D},\ Z_i = 1\},\ I = \{x_i \mid (x_i, y_i^\star) \in \mathcal{D},\ Z_i = 0\} \tag{4}$$

For each sample x_i, let $B_i \in \{L, H\}$ denote its entropy band, obtained by thresholding its semantic entropy (Eqn 1). We define four entropy-based subsets: $C_L = \{x_i \in C \mid B_i = L\}$, $C_H = \{x_i \in C \mid B_i = H\}$, $I_L = \{x_i \in I \mid B_i = L\}$, $I_H = \{x_i \in I \mid B_i = H\}$. Let $\hat{Z}(x) \in [0, 1]$ denote the correctness score predicted by the probe and $\mathrm{AUC}(C_a, I_b)$ is the probability that a randomly drawn correct sample from entropy band a receives a higher score than an incorrect sample from band b, for $a, b \in \{L, H\}$. Motivated by the xAUC measure Kallus & Zhou (2019) to detect shortcut, we evaluate both within-band and cross-band correctness prediction performance. We define ES-xAUC as the worst performance across the four atomic comparisons:

$$\text{ES-xAUC} = \min\Big\{\mathrm{AUC}(C_H, I_H), \mathrm{AUC}(C_H, I_L), \mathrm{AUC}(C_L, I_H), \mathrm{AUC}(C_L, I_L)\Big\} \tag{5}$$

A higher ES-xAUC therefore indicates stronger worst-case correctness prediction performance. We measure the variation among these entropy subset based comparisons using:

$$\text{ES-}\delta\text{AUC} = \max\Big\{\mathrm{AUC}(C_H, I_H), \mathrm{AUC}(C_H, I_L), \mathrm{AUC}(C_L, I_H), \mathrm{AUC}(C_L, I_L)\Big\} - \text{ES-xAUC} \tag{6}$$

A lower ES-δAUC indicates uniform correctness prediction performance. Finally, we report prediction performance in the confidence-misaligned direction, $\mathrm{AUC}(C_H, I_L)$, which compares high entropy correct samples against low entropy incorrect samples. We formally establish that these metrics help quantify entropy-shortcut in Appendix A.2.

3.2 Tasks and models

Datasets: We evaluate open-ended QA across text and vision benchmarks **totaling $\sim 165k$ samples** across eight datasets with four LLMs and five multimodal. For *text QA*, we use SciQ Welbl et al. (2017), TriviaQA Joshi et al. (2017), and Natural Questions (NQ) Kwiatkowski et al. (2019). For *visual QA*, we use OK-VQA Marino et al. (2019), VQAv2 Antol et al. (2015), and AdVQA Sheng

Table 1: **Benchmarking entropy-misaligned predictions.** The table below shows the prediction accuracies of the LLMs and multimodal LLMs along with the proportion of high-entropy correct predictions and low-entropy mispredictions using a large set of samples (Appendix Table 6).

Text QA															
	Natural Questions					SciQ					TriviaQA				
Statistic	**Gemma**	**Llama**	**Mistral**	**Qwen**	**Avg.**	**Gemma**	**Llama**	**Mistral**	**Qwen**	**Avg.**	**Gemma**	**Llama**	**Mistral**	**Qwen**	**Avg.**
Accuracy (%)	47.8	50.8	52.3	40.9	48.0	87.7	87.0	83.4	87.5	86.4	73.2	77.2	75.6	65.4	72.9
$\mathbf{C_H}$ **(% of C)**	8.8	41.9	40.1	24.9	28.9	8.2	23.9	29.2	20.0	20.3	5.4	19.8	13.7	8.0	11.8
$\mathbf{I_L}$ **(% of I)**	62.6	18.8	23.0	31.6	34.0	68.0	31.8	32.7	44.4	44.2	55.7	18.0	30.7	42.4	36.7

Image QA												
	AdvQA				OK-VQA				VQAv2			
Statistic	**Gemma**	**Pixtral**	**QwenV**	**Avg.**	**Gemma**	**Pixtral**	**QwenV**	**Avg.**	**Gemma**	**Pixtral**	**QwenV**	**Avg.**
Accuracy (%)	53.5	47.5	58.6	53.2	64.5	68.3	70.8	67.9	65.3	69.6	75.0	70.0
$\mathbf{C_H}$ **(% of C)**	12.8	21.9	32.4	22.3	17.7	20.6	45.0	27.8	15.6	33.4	33.2	27.4
$\mathbf{I_L}$ **(% of I)**	62.2	44.7	36.1	47.7	61.6	45.3	18.8	41.9	57.5	41.6	24.9	41.3

The Table 1 uses shortened notations of `Gemma-3-12B`, `Llama-3.1-8B`, `Mistral-7B-v0.3`, `Qwen2.5-7B`, `Pixtral-12B`, and `Qwen2.5-VL-7B`.

et al. (2021). We additionally consider two safety-critical medical image-classification tasks: three-class pneumonia classification on the RSNA Pneumonia Detection Challenge Wang et al. (2017) and binary tumour detection on Camelyon17-WILDS Bandi et al. (2019). Dataset sizes, retained training counts, and held-out evaluation sizes are reported in Appendix A.1, Tables 6 and 17.

Models: We evaluate four open-weight LLMs: Llama-3.1-8B-Instruct Grattafiori et al. (2024), Mistral-7B-Instruct-v0.3 Jiang et al. (2023), Qwen2.5-7B-Instruct Qwen et al. (2025), and Gemma-3-12B-it Team et al. (2025). Our VLM evaluation uses Qwen2.5-VL-7B-Instruct Bai et al. (2025), Gemma-3-12B-it Team et al. (2025), and Pixtral-12B Agrawal et al. (2024). For the medical tasks, we fine-tune Qwen2-VL-2B-Instruct Bai et al. (2025) and SmolVLM2-2.2B-Instruct Marafioti et al. (2025).

Sampling entropy and bands: For each input, we compute semantic entropy Kuhn et al. (2023) from 10 responses sampled at $T = 1.0$. We partition samples into low-entropy (L) and high-entropy (H) bands using the two-class semantic-entropy binarization procedure of Han et al. (2024). Further details are provided in Appendices A.8 and A.8.1.

We use `Llama-3.3-70B-Instruct` Grattafiori et al. (2024) as the answer judge. The judge prompt and probe-training details are provided in Appendices A.3 and A.4.

4 Benchmarking correctness-entropy associations

We perform extensive benchmarking totaling $\sim 165k$ samples across eight datasets with four LLMs and five multimodal LLMs to understand the proportion of entropy-misalignment among the correct and inaccurate samples, based on the entropy-based subsets (CH, CL, IH, IL), defined in Section 3. The results are shown in Table 1. We observe that C_H and I_L form a considerable proportion of the samples in C and I, respectively. We observe $I_L(\%\text{ of } I)$ to be significantly more than $C_H(\%\text{ of } C)$, pointing to a larger set of low-entropy mis-predictions (certain mis-predictions). Moreover, such unusual associations become even more abundant with multimodal LLMs. For LLMs, high-entropy correct predictions and low-entropy mis-predictions average 20.3% and 38.3% of the correct and mis-predicted samples, respectively. For multimodal LLMs, the same measures are 25.8% and 43.6%.

5 Do correctness probes use entropy-shortcut?

5.1 Entropy subset analysis

Given the evidence of entropy-correctness dissociation, we move to analyze the cross-group performances between entropy-based subsets. The results are shown in Table 2. We make interesting observations,

Table 2: **Dataset-wise correctness prediction AUC** of semantic entropy Kuhn et al. (2023), semantic entropy probe (SEP) Han et al. (2024), and the correctness probe Orgad et al. (2025), dissected across correctness-entropy subsets. The results are averaged over four LLMs for text QA tasks. Model-wise results are reported in Appendix Table 7.

Dataset	Semantic Entropy						SEP						Correctness Probe					
	CvI	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC	CvI	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC	CvI	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC
Text QA																		
NQ	0.750	0.651	0.000	1.000	0.634	0.000	0.754	0.653	0.280	0.928	0.689	0.280	0.839	0.807	0.592	0.924	0.791	0.592
SciQ	0.728	0.618	0.000	1.000	0.624	0.000	0.777	0.646	0.394	0.908	0.748	0.394	0.852	0.782	0.643	0.921	0.838	0.643
TriviaQA	0.830	0.709	0.000	1.000	0.690	0.000	0.875	0.767	0.484	0.949	0.818	0.484	0.905	0.829	0.630	0.956	0.867	0.630
Image QA																		
VQAv2	0.700	0.555	0.000	1.000	0.643	0.000	0.769	0.678	0.339	0.928	0.727	0.339	0.845	0.801	0.588	0.935	0.824	0.588
OK-VQA	0.707	0.598	0.000	1.000	0.635	0.000	0.750	0.641	0.258	0.940	0.708	0.258	0.834	0.790	0.564	0.932	0.805	0.564
AdvQA	0.707	0.616	0.000	1.000	0.636	0.000	0.723	0.658	0.265	0.910	0.661	0.265	0.778	0.696	0.502	0.882	0.765	0.502

- **Semantic entropy:** Semantic entropy itself is used to predict correctness for LLM predictions in black-box settings Kuhn et al. (2023); Farquhar et al. (2024), as it offers somewhat high correctness prediction AUC (CvI). However, violations of the correctness-entropy inverse relationship invalidate this, making it the most direct example of an entropy shortcut which collapses the CHvIL and ES-xAUC, as reflected in Table 2. The high values of CvI AUC come from exploiting the usual correctness-entropy association.
- **Semantic entropy probe (SEP):** In SEP, the LLM hidden states are fed to a linear model to predict semantic entropy. A logistic regression model takes the inference time hidden state as input and is trained to predict high (H)/low (L) semantic entropy bands. At test time, SEP achieves a superior correctness-detection AUC Han et al. (2024). As evident in Table 2, the gains are obtained by exploiting entropy-shortcut (CHvIL AUC and ES-xAUC remain well below 50%).
- **Correctness probe:** The correctness probes are exactly like SEP by construction; however, they are trained using correctness labels instead of semantic entropy. The improvements here in CHvIL AUC further confirm that the probe uses an entropy shortcut when semantic entropy is used for training (SEP). The ES-xAUC also comes above 50%. However, the large gap in subset AUCs points to the broader problem that even with correctness labels, probes inherently pick up the entropy-shortcut.

Summary: Across the three common predictors, including the LLM correctness probe, the worst-case atomic correctness discrimination, measured by ES-xAUC, remains below chance or only slightly above it. In contrast, high pooled correct-versus-incorrect AUROC (CvI) can obscure poor performance on entropy-misaligned subsets.

5.2 Controlled entropy analysis

`Matched` training: Given the indications of an entropy shortcut in the LLM correctness probe, we seek further evidences behind it. Based on Table 1, samples in $\{CL, IH\}$ constitute 75.2% of the total samples $\{C, I\}$ in text QA and 66.7% in image QA tasks. Correctness prediction for $\{CL, IH\}$ can be perfected by taking an entropy-shortcut. We perform a controlled study in which the shortcut attribute of entropy is controlled in the training dataset. We perform entropy matching between the correct (C) and incorrect (I) samples using constrained sampling from both sets. We call this modified training set the `Matched` train split. This removes entropy as an easy discriminating signal, and the LLM probe is then trained on this matched-entropy dataset. The test split remains unchanged. The entropy-matching steps are provided in Appendix A.7.1.

The prediction AUC and subset AUCs are shown in Table 3. Three metrics- CHvIL, ES-xAUC and ES-δAUC are also shown and compared against the original `Natural` training split. We observe significant improvements in shortcut mitigation, with average ES-xAUC (and CHvIL AUC) increasing from 62.2% to 71.7% for text QA and from 55.1% to 66.7% for image QA. Moreover, ES-δAUC also reduces from 31.2% to 14.9% for text QA and from 36.5% to 15.8% for image QA. However, it is interesting to note that an LLM probe trained on correctness labels still exhibits entropy-shortcut behavior to a considerable degree even after entropy-matching, as evidenced by $\sim 15\%$ ES-δAUC across text and image QA tasks.

Table 3: **Matched-entropy training of LLM correctness probe.** The correctness prediction AUC of the LLM probe when trained using `Matched` training split, averaged across LLMs, and compared against the original `Natural` train split. Detailed results are shown in Appendix Table 12.

Dataset	Trained on: Natural							Trained on: Matched						
	CvI	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC	ES-δAUC	CvI	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC	ES-δAUC
Text QA														
NQ	0.839	0.807	0.592	0.924	0.791	0.592	0.332	0.801	0.797	0.733	0.830	0.771	0.733	0.097
SciQ	0.852	0.782	0.643	0.921	0.838	0.643	0.279	0.821	0.776	0.711	0.860	0.815	0.711	0.149
TriviaQA	0.905	0.829	0.630	0.956	0.867	0.630	0.326	0.872	0.806	0.707	0.908	0.844	0.707	0.202
Image QA														
VQAv2	0.845	0.801	0.588	0.935	0.824	0.588	0.347	0.819	0.797	0.709	0.867	0.809	0.709	0.158
OK-VQA	0.834	0.790	0.564	0.932	0.805	0.564	0.368	0.800	0.776	0.698	0.845	0.785	0.698	0.147
AdvQA	0.778	0.696	0.502	0.882	0.765	0.502	0.380	0.718	0.638	0.594	0.764	0.732	0.594	0.170

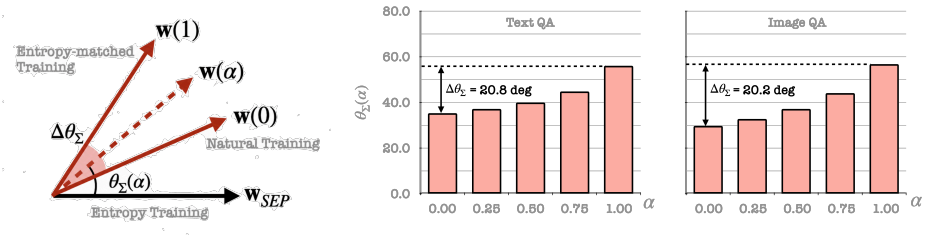


Figure 2: **Geometric interpretation of entropy-shortcut**: An LLM probe trained on gradual entropy-matching makes the probe weight monotonically move away from the entropy direction represented by SEP, for $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$. Detailed results are in Appendix Table 13.

Geometric interpretation: To obtain geometric evidence of shortcut learning, we compare LLM correctness probes learned using the `Natural` and `Matched` training splits. Let Δ denote the difference in mean entropy between correct and incorrect samples in the `Natural` split. We construct intermediate datasets $\mathcal{D}(\alpha)$, with $\alpha \in [0, 1]$, such that this difference is reduced to $(1-\alpha)\Delta$. Thus, $\mathcal{D}(0)$ corresponds to `Natural`, while $\mathcal{D}(1)$ corresponds to `Matched`, where the entropy difference is zero. An LLM correctness probe trained on $\mathcal{D}(\alpha)$ is given by

$$\hat{Z} = \sigma\left(\mathbf{w}_\alpha^\top h_\mathcal{M}(X) + \beta_\alpha\right). \tag{7}$$

Also, the SEP, an LLM probe that is explicitly trained on semantic entropy, can be expressed as, $\widehat{SE}(X) = \sigma\left(\mathbf{w}_{SEP}^\top h_\mathcal{M}(X) + \beta_{SEP}\right)$. We measure the angle between $\mathbf{w}_\alpha$ and $\mathbf{w}_{SEP}$, and a lesser value signifies stronger entropy-shortcut. Because of the large dimension of LLM hidden states, the standard cosine angle is not appropriate, and we use the covariance-weighted cosine angle instead Ying et al. (2026a;b).

$$\theta_\Sigma(\alpha) = \cos^{-1}\left[\frac{{\mathbf{w}_{SEP}}^\top \widehat{\Sigma} \mathbf{w}_\alpha}{\sqrt{\mathbf{w}_{SEP}^\top \widehat{\Sigma} \mathbf{w}_{SEP}}\sqrt{\mathbf{w}_\alpha^\top \widehat{\Sigma} \mathbf{w}_\alpha}}\right] \tag{8}$$

where $\widehat{\Sigma}$ is the empirical covariance matrix of centered hidden states $h_\mathcal{M}(X)$. Figure 2 shows the $\theta_\Sigma(\alpha)$ averaged across LLMs and datasets. For the values of $\alpha = \{0, 0.25, 0.5, 0.75, 1.0\}$, $\theta_\Sigma(\alpha)$ increases with α. This controlled experiment shows the gradual diminishing of the entropy shortcut as the correctness probe weight moves away from the SEP. $\Delta\theta_\Sigma = \theta_\Sigma(1) - \theta_\Sigma(0)$ achieves significant deviation of 20.5 degrees. **Summary:** Matching entropy between correct and incorrect samples while training an LLM probe partially alleviates the entropy shortcut. The shortcut diminishes as the training dataset approaches `Matched` from `Natural`. However, the entropy-shortcut persists considerably even after `Matched` training.

6 Mitigating Entropy-Shortcut

6.1 Joint Learning

In this analysis, we ask the question - under a modified training objective, can we learn the same from the `Natural` training split itself? To do this, we separately train on entropy subsets where the

optimization in each subset predicts correctness. We first decompose the binary cross-entropy loss into four entropy subsets.

$$\mathcal{L}_G(\mathbf{w}, \beta) = \frac{1}{|G|} \sum_{x_i \in G} \left[-Z_i \log \widehat{Z}(x_i) - (1 - Z_i) \log \left(1 - \widehat{Z}(x_i) \right) \right], \quad G \in \{C_H, C_L, I_H, I_L\} \tag{9}$$

Then we keep an objective that encourages the probe to perform consistently across the entropy subsets. We take two different optimization strategies to achieve the same.

Uniform weighting: In the first approach, to weigh the correctness prediction performance equally across entropy subsets, we optimise a joint BCE loss with equal weightage for each subset. The objective to update the correctness probe parameters becomes (with $\lambda = 1.0$ weight regularization):

$$\mathcal{L}_{\text{uniform}}(\mathbf{w}, \beta, \mathbf{q}) = \sum_{G \in \{C_H, C_L, I_H, I_L\}} q_G \mathcal{L}_G(\mathbf{w}, \beta) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2, \quad q_G = \frac{1}{4} \tag{10}$$

Worst group optimization: Instead of static uniform weighting, we attempt non-uniform weighting, in which the worst-performing group receives more weight dynamically based on the distributionally robust optimization (DRO) algorithm Sagawa* et al. (2020). We denote this objective as $\mathcal{L}_{\text{DRO}}(\mathbf{w}, b, \mathbf{q})$ where the group weights (q_G) are dynamically updated at each iteration. The q_G's start with uniform weighting, and subsequently give more weight to the worst group.

$$q_G \leftarrow \frac{q_G \exp(\mathcal{L}_G(\mathbf{w}, \beta))}{\sum_{G' \in \{C_H, C_L, I_H, I_L\}} q_{G'} \exp(\mathcal{L}_{G'}(\mathbf{w}, \beta))}, \qquad q_G^{(0)} = \frac{1}{4} \tag{11}$$

As shown in Table 4, joint learning helps significantly reduce the entropy-shortcut with equivalent results to those of `Matched` training, although trained on the `Natural` split. They are also compared against other LLM probes Binkowski et al. (2025); Sriramanan et al. (2024); Binkowski et al. (2026). Table 4 reports worst group optimization results which performs better than uniform weighting. The results with uniform weighting are reported in the Appendix Table 8. Compared to the best prior works, joint learning improves in all metrics- ES-xAUC and ES-δAUC improve by 9.8% and 14.5%, respectively, for text QA, and 11.9% and 5.8% for image QA as seen in Table 14.

6.2 Routed learning

Since `Matched` training and entropy-subset-based joint learning only partially mitigate the entropy shortcut, we next avoid attempting to mitigate the entropy-shortcut across all entropy bands using a single probe. As illustrated in Figure 3, a router first assigns each sample to the high-entropy (H) or low-entropy (L) band. We then train two separate probes: an H-band probe on $C_H \cup I_H$ and an L-band probe on $C_L \cup I_L$ for predicting correctness within those bands. We refer to this approach as *routed learning*. Assuming the entropy-independent correctness signal remains stable across the two entropy bands, two probes learn the shared correctness signal within their bands, while the associations between entropy and correctness in each band are absorbed by their respective intercepts. Since low-entropy predictions are more frequently correct, this predicts $\beta_L > \beta_H$. We observe this ordering for every model-dataset pair; moreover, the band-specific probes do not meaningfully outperform the general correctness probe within their own bands. This evidence suggests that routing isolates band-dependent offsets rather than producing stronger within-band classifiers. We perform z-normalization on each probe scores and the normalized scores are pooled to obtain the final correctness score. This removes the biased entropy-associations and hence, the shortcut. We formalize this argument and report the supporting intercept and within-band results in Appendix A.6.

- **Training the router:** On the train set, we train a probe to predict the entropy labels, similar to SEP. They take the sampling hidden states of the LLM and predict H vs L. The trained probe is directly used for the held-out test set. In inference time, the router classifies every sample to either belong to H or L with 95.4% AUC on the test split. The router training details and the routing performance are reported in A.6.
- **Separate LLM probes:** Two separate LLM correctness probes are trained. They are trained to predict correctness within samples belonging to high (H) and low (L) entropy bands. Their correctness prediction AUC indeed remains similar, and their intercepts significantly vary, satisfying $\beta_L > \beta_H$ across all datasets and LLMs (Figure 3). Further details are provided in A.6.4.

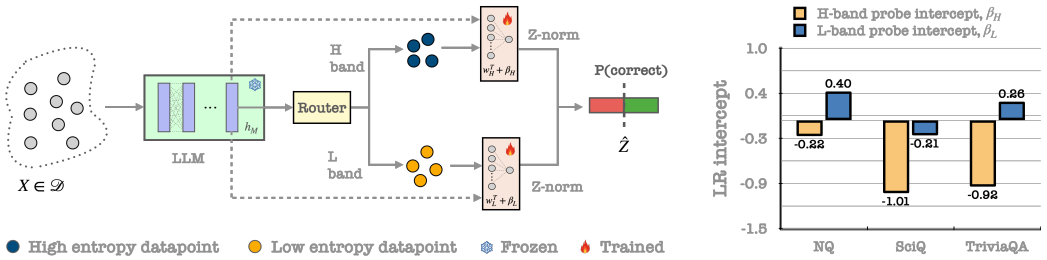


Figure 3: **Routed learning.** A learned router assigns each sample to the high- or low-entropy band. The corresponding band-specific correctness probe produces a logit, which is normalized using that probe's training-set statistics before the two score streams are pooled.

Table 4: **Correctness prediction performance of the proposed LLM probe.** The proposed joint learning and routed learning probes are compared against several prior works. All evaluated prior LLM probes exhibit evidence of the entropy shortcut. The model-wise results are in Appendix Table 15.

Text QA										
Method	**Band AUC**	**AUC(C_H, I_L) ↑**			**ES-xAUC ↑**			**ES-δAUC ↓**		
		NQ	**SciQ**	**TriviaQA**	**NQ**	**SciQ**	**TriviaQA**	**NQ**	**SciQ**	**TriviaQA**
Semantic Entropy		0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000
SEP		0.280	0.394	0.484	0.280	0.394	0.484	0.648	0.514	0.466
Correctness Probe		0.592	0.643	0.630	0.592	0.643	0.630	0.332	0.279	0.326
SINK		0.556	0.597	0.567	0.556	0.597	0.567	0.348	0.294	0.379
LapEigVals		0.534	0.516	0.526	0.534	0.516	0.526	0.351	0.325	0.386
AttnLogDet		0.552	0.569	0.562	0.552	0.569	0.562	0.341	0.215	0.342
AttnEigVals		0.541	0.543	0.544	0.541	0.543	0.544	0.327	0.237	0.322
Joint Learning (DRO)		0.731	0.720	0.709	0.731	0.720	0.709	0.110	0.140	0.202
Routed Learning		**0.804**	**0.800**	**0.846**	**0.737**	**0.749**	**0.723**	**0.068**	**0.052**	**0.125**

Band-AUC order: The order of the cells in band AUC is {CHvIH,CHvIL,CLvIH,CLvIL}. Colors indicate AUROC: 0.0 1.0

- **Z-normalization and pooling:** The correctness scores obtained from each of the probes is first z-normalized and the pooled together. This normalizes away the band-specific score offsets captured by the expert intercepts and thereby reduces the entropy shortcut. During test time, z-normalization is done using normalization parameters learnt on the train set.

The results are shown in Table 4. The routed learning significantly boosts all three metrics, further disentangling the entropy shortcut. It achieves the best performance across all three metrics, providing significant gains over joint learning and `Matched` training. The AUC(C_H, I_L), ES-xAUC and ES-δAUC improve by 9.7%, 1.6% and 6.9% over the joint learning. This robustness comes with a lower pooled AUC. However, the proposed approach only intended to measure the shortcut-removed correctness prediction performance. Detailed results with multimodal LLMs are provided in Appendix Table 14 and 15.

Summary: An LLM correctness probe can be trained with reduced entropy-shortcut using entropy subset-aware joint-optimization. Moreover, a routed training approach with two separate probes trained on high and low-entropy samples followed by Z-normalization reduces the entropy shortcut.

7 Discussion

Effect of language prior: In the case of multimodal LLMs, additional non-text modalities bring in other failure modes. A major one is the language prior (LP) where the model heavily relies on the textual component of input and ignores the non-text modality. To obtain language prior labels, we exploit complementary image pairs in VQAv2 Antol et al. (2015) and label an input as LP when it produces the same answer for both images and exactly one of them is correct. Interestingly, we find that correctness-entropy composition changes uniquely when LP is present. As shown in Figure 4(a), language-prior reliance shifts both correct and incorrect predictions toward the low-entropy band. In particular, it substantially enhances low-entropy mispredictions.

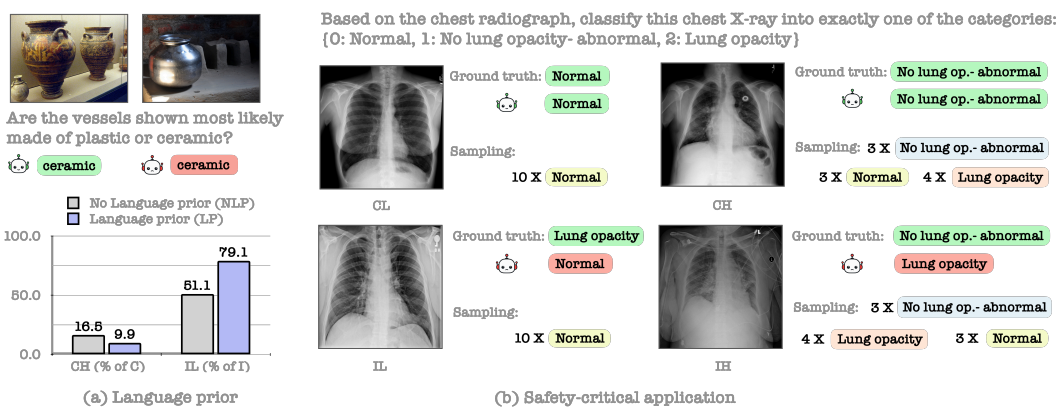


Figure 4: **(a) Effect of language prior:** In multimodal QA, language prior (LP) alters the entropy-correctness composition. Occurrence of LP shifts all samples more towards the low-entropy band. **(b) Safety-critical application:** Entropy-correctness misalignment continues to hold in fine-tuned LLMs for medical applications, making the entropy-shortcut critically vulnerable.

Safety-critical applications: We further evaluate the effect of entropy-shortcut in critical medical applications. We take Camelyon17 Bandi et al. (2019) that requires detecting tumor tissue in histopathology images, and RSNA Shih et al. (2019) for classifying chest radiographs as three classes-normal, abnormal without lung opacity, or lung opacity. We fine-tune Qwen2-VL-2B and SmolVLM2-2.2B on each task. We observe significant proportions of C_H and I_L using them, and examples of all four associations are shown in Figure 4(b). As shown in Appendix Table 18, routed learning increases $\mathrm{AUC}(C_H, I_L)$ for Camelyon17 to 0.862, and on RSNA, it improves both $\mathrm{AUC}(C_H, I_L)$ and ES-xAUC.

8 Prior Works

Post-hoc correctness detectors use either sampled outputs or model internals. Black-box methods use sampling to quantify disagreement among stochastic generations like SelfCheckGPT Manakul et al. (2023), semantic entropy Kuhn et al. (2023), and kernel language entropy Nikitin et al. (2024). White-box methods instead predict semantic entropy Han et al. (2024) or correctness Orgad et al. (2025) from hidden states, or use attention-derived signals like LLM-Check Sriramanan et al. (2024), Lookback Lens Chuang et al. (2024), and spectral attention features Binkowski et al. (2025). However, recent works point out the existence of low-entropy mispredictions and that consistency is not equivalent to correctness. They show that multiple detector families deteriorate on such errors Simhi et al. (2025); Gekhman et al. (2025); Tan et al. (2025). These studies establish this as a failure mode, but neither formalize it nor examine whether correctness probes themselves exploit entropy as a shortcut. Empirical risk minimization for correctness detection can achieve strong aggregate performance while failing on shortcut-conflicting groups as shown in prior shortcut learning literature Sagawa* et al. (2020); Liu et al. (2021). From this perspective, for entropy-correctness groups, we develop joint and routed learning probes that remain robust across entropy-based subgroups.

9 Conclusion

We identify the entropy-shortcut as a systematic failure mode of post-hoc LLM correctness probes. Across four LLMs and five VLMs evaluated on eight datasets, high-entropy correct predictions and low-entropy incorrect predictions constitute substantial fractions of model inferences. Although standard correctness probes are trained only on correctness labels and do not explicitly receive entropy labels, they inherit the entropy-correctness association in the natural training distribution. Consequently, a strong aggregate AUROC can conceal substantially weaker discrimination between high entropy correct and low entropy incorrect predictions. Our entropy-conditioned subset analysis, matched-entropy training, and geometric interpretations make this failure evident. We also propose joint learning and routed learning approaches to improve worst-group discrimination and consistency across all entropy bands.

Reproducibility statement

The appendices provide the models, datasets, prompts, data splits, hyperparameters, algorithms, and model-dataset pair level results required to reproduce our experiments. The implementation codes will be made publicly available upon acceptance.

AI use statement

Generative AI tools were used to assist with literature discovery and summarization, critique and refinement of the conceptual and theoretical presentation and experimental analyses, manuscript editing, and table and figure preparation. They were not used to generate experimental measurements. All AI-assisted suggestions, derivations, and comments were independently checked by the authors, who take full responsibility for the final work.

Ethics statement

We use publicly available benchmark datasets and open-weight models and collect no new human-subject data. The medical experiments are research-only and should not be interpreted as clinical validation or deployment guidance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2019. ISSN 1558-254X. doi: 10.1109/TMI.2018.2867350.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 333(4):319–328, 2025.

Debarpan Bhattacharya, Malay Phadke, and Sriram Ganapathy. Big-sure-bipartite graph for semantic uncertainty and reliability estimation of llms. *arXiv preprint arXiv:2608.30646*, 2026.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Hallucination detection in llms using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24354–24385, 2025.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3NdFFpcEjU.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Walter D Fisher. On grouping for maximum homogeneity. *Journal of the American statistical Association*, 53(284):789–798, 1958.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299*, 2025.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International conference on machine learning*, pp. 6781–6792. PMLR, 2021.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3190–3199. IEEE, 2019.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. *arXiv preprint arXiv:2506.00245*, 2025.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In *Proceedings on*, pp. 49–64. PMLR, 2023.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, i'm wrong: High-certainty hallucinations in llms. *arXiv preprint arXiv:2502.12964*, 2025.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in neural information processing systems*, 37:34188–34216, 2024.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, et al. Too consistent to detect: A study of self-consistent errors in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18128–18138, 2024.

Zhuofan Josh Ying, Peter Hase, and Nikolaus Kriegeskorte. Comparing linear probes with mahalanobis cosine similarity. *arXiv preprint arXiv:2606.19603*, 2026a.

Zhuofan Josh Ying, Shauli Ravfogel, Nikolaus Kriegeskorte, and Peter Hase. The truthfulness spectrum hypothesis. *arXiv preprint arXiv:2602.20273*, 2026b.

A Appendix

A.1 Benchmarking correctness-entropy associations

A.1.1 Train/test split statistics

Table 5 summarizes the number of examples processed during inference and subsequently used to train and evaluate the correctness probes. For TriviaQA, NQ, and VQAv2, inference is performed on a larger pool, from which $15{,}000$ examples are subsampled for probe experiments. SciQ, AdvQA, and OK-VQA use all available inference examples. All text- and image-QA probe splits use a $70/30$ train/test partition stratified jointly by correctness and entropy band, thereby preserving the four atomic subsets C_L, C_H, I_L, and I_H.

For the medical tasks, correctness-probe performance is averaged over five independently seeded $60/40$ splits, each stratified by the four atomic subsets. These probe splits are distinct from the data used to fine-tune the generative VLMs. RSNA contains $26{,}684$ images: $21{,}348$ training, $2{,}668$ validation, and $2{,}668$ test images. The VLMs are fine-tuned using the official training split and evaluated for correctness detection on the $2{,}668$-image test split. For Camelyon17, the VLMs are fine-tuned on the $50{,}000$-patch source-hospital training set, while correctness detection is evaluated on the $20{,}000$-image out-of-distribution validation split.

Table 5: Inference and correctness-probe split sizes. Row counts are reported per model. The medical rows correspond to the held-out images used for correctness-probe evaluation, not to the data used for VLM fine-tuning. For the medical tasks, the probe train/test sizes are per split: each of five random, stratified 60/40 splits uses the sizes shown, and results are averaged over the five splits.

Dataset	Inference rows	Probe rows	Probe train/test	Notes
TriviaQA	50,000	15,000	10,500 / 4,500	Subsampled after inference.
NQ	50,000	15,000	10,500 / 4,500	Subsampled after inference.
SciQ	13,564	13,564	9,495 / 4,069	All available examples.
AdvQA	3,000	3,000	2,100 / 900	All available examples.
OK-VQA	14,055	14,055	$\approx 9{,}839/4{,}216$	All available examples.
VQAv2	34,991	15,000	10,500 / 4,500	Subsampled after inference.
RSNA	2,668	2,668	$\approx 1{,}600/1{,}068$	Official held-out test split; 5 random 60/40 splits.
Camelyon17	20,000	20,000	$\approx 11{,}996/8{,}004$	OOD validation split; 5 random 60/40 splits.

A.1.2 Model–dataset-level correctness–entropy composition

Table 6 expands the model-averaged results reported in the main paper by presenting the complete correctness–entropy composition for every model–dataset pair. Across all 21 pairs, both entropy-misaligned subsets, C_H and I_L, contain substantial numbers of samples. Thus, the averaged trends are not driven by any individual model or dataset; violations of the conventional correctness–entropy association occur consistently across both text and image QA.

- **Misalignment occurs in every setting.** The proportion of C_H ranges from 5.4% to 41.9% in text QA and from 12.8% to 45.0% in image QA. Similarly, the proportion of I_L ranges from 18.0% to 68.0% in text QA and from 18.8% to 62.2% in image QA. Hence, neither high-entropy correct predictions nor low-entropy incorrect predictions are isolated edge cases.
- **Low-entropy mispredictions are particularly prevalent.** In 16 of the 21 model–dataset pairs, I_L constitutes a larger fraction of incorrect predictions than C_H constitutes of correct predictions. Averaged across datasets and models, I_L accounts for 38.3% of text-QA errors and 43.6% of image-QA errors, compared with 20.3% and 25.8% for C_H, respectively.
- **The composition is strongly model-dependent.** For example, Gemma consistently produces relatively few C_H predictions but many I_L predictions. Across the six QA datasets, its C_H proportion ranges from 5.4% to 17.7%, whereas its I_L proportion ranges from 55.7% to 68.0%. In contrast, Qwen2.5-VL exhibits substantially larger C_H and smaller I_L proportions on image

Table 6: Detailed dataset statistics and atomic correctness–entropy subsets across text and image QA benchmarks. Each entry reports the sample count, followed by the percentage in parentheses. Percentages for C and I are computed relative to N; percentages for C_L and C_H are computed relative to C; and percentages for I_L and I_H are computed relative to I.

Text QA												
	Natural Questions				**SciQ**				**TriviaQA**			
Statistic	**Gemma**	**Llama**	**Mistral**	**Qwen**	**Gemma**	**Llama**	**Mistral**	**Qwen**	**Gemma**	**Llama**	**Mistral**	**Qwen**
N	15,000	15,000	15,000	15,000	13,564	13,564	13,564	13,564	15,000	15,000	15,000	15,000
C	7,172 (47.8%)	7,624 (50.8%)	7,843 (52.3%)	6,139 (40.9%)	11,896 (87.7%)	11,797 (87.0%)	11,312 (83.4%)	11,864 (87.5%)	10,981 (73.2%)	11,576 (77.2%)	11,340 (75.6%)	9,815 (65.4%)
I	7,828 (52.2%)	7,376 (49.2%)	7,157 (47.7%)	8,861 (59.1%)	1,668 (12.3%)	1,767 (13.0%)	2,252 (16.6%)	1,700 (12.5%)	4,019 (26.8%)	3,424 (22.8%)	3,660 (24.4%)	5,185 (34.6%)
C_L	6,541 (91.2%)	4,427 (58.1%)	4,695 (59.9%)	4,610 (75.1%)	10,918 (91.8%)	8,972 (76.1%)	8,011 (70.8%)	9,490 (80.0%)	10,384 (94.6%)	9,282 (80.2%)	9,782 (86.3%)	9,025 (92.0%)
C_H	631 (8.8%)	3,197 (41.9%)	3,148 (40.1%)	1,529 (24.9%)	978 (8.2%)	2,825 (23.9%)	3,301 (29.2%)	2,374 (20.0%)	597 (5.4%)	2,294 (19.8%)	1,558 (13.7%)	790 (8.0%)
I_L	4,897 (62.6%)	1,384 (18.8%)	1,648 (23.0%)	2,798 (31.6%)	1,134 (68.0%)	562 (31.8%)	737 (32.7%)	755 (44.4%)	2,237 (55.7%)	618 (18.0%)	1,123 (30.7%)	2,199 (42.4%)
I_H	2,931 (37.4%)	5,992 (81.2%)	5,509 (77.0%)	6,063 (68.4%)	534 (32.0%)	1,205 (68.2%)	1,515 (67.3%)	945 (55.6%)	1,782 (44.3%)	2,806 (82.0%)	2,537 (69.3%)	2,986 (57.6%)

Image QA									
	AdvQA			**OK-VQA**			**VQAv2**		
Statistic	**Gemma**	**Pixtral**	**QwenV**	**Gemma**	**Pixtral**	**QwenV**	**Gemma**	**Pixtral**	**QwenV**
N	3,000	3,000	3,000	14,055	14,055	14,055	15,000	15,000	15,000
C	1,605 (53.5%)	1,425 (47.5%)	1,758 (58.6%)	9,061 (64.5%)	9,593 (68.3%)	9,946 (70.8%)	9,792 (65.3%)	10,433 (69.6%)	11,248 (75.0%)
I	1,395 (46.5%)	1,575 (52.5%)	1,242 (41.4%)	4,994 (35.5%)	4,462 (31.7%)	4,109 (29.2%)	5,208 (34.7%)	4,567 (30.4%)	3,752 (25.0%)
C_L	1,400 (87.2%)	1,113 (78.1%)	1,189 (67.6%)	7,453 (82.3%)	7,617 (79.4%)	5,474 (55.0%)	8,262 (84.4%)	6,949 (66.6%)	7,517 (66.8%)
C_H	205 (12.8%)	312 (21.9%)	569 (32.4%)	1,608 (17.7%)	1,976 (20.6%)	4,472 (45.0%)	1,530 (15.6%)	3,484 (33.4%)	3,731 (33.2%)
I_L	868 (62.2%)	704 (44.7%)	448 (36.1%)	3,077 (61.6%)	2,023 (45.3%)	774 (18.8%)	2,994 (57.5%)	1,901 (41.6%)	934 (24.9%)
I_H	527 (37.8%)	871 (55.3%)	794 (63.9%)	1,917 (38.4%)	2,439 (54.7%)	3,335 (81.2%)	2,214 (42.5%)	2,666 (58.4%)	2,818 (75.1%)

C_L and C_H denote correct low- and high-entropy predictions, respectively; I_L and I_H denote incorrect low- and high-entropy predictions. The shaded rows, C_H and I_L, are the entropy-misaligned subsets. **Gemma**, **Llama**, **Mistral**, **Qwen**, **Pixtral**, and **QwenV** denote `Gemma-3-12B`, `Llama-3.1-8B`, `Mistral-7B-v0.3`, `Qwen2.5-7B`, `Pixtral-12B`, and `Qwen2.5-VL-7B`, respectively.

QA, reaching 45.0% C_H and only 18.8% I_L on OK-VQA. Sampling entropy therefore reflects model-specific response behaviour rather than a dataset property alone.

- **Higher task accuracy does not eliminate entropy misalignment.** On SciQ, all four LLMs achieve between 83.4% and 87.7% accuracy, yet 31.8%–68.0% of their remaining errors belong to I_L. Thus, strong task performance can coexist with a large population of consistently reproduced errors.
- **Aggregate evaluation can conceal the difficult subsets.** The prevalence of the entropy-aligned subsets, C_L and I_H, makes entropy a useful shortcut on the natural distribution. However, the consistently non-negligible C_H and I_L populations show why aggregate correctness AUROC alone is insufficient and motivate the entropy-conditioned evaluation used in the main paper.

A.2 Properties of the entropy-shortcut metrics

The main paper introduces entropy-conditioned atomic AUROCs, ES-xAUC, and ES-δAUC to expose variation in correctness detection across entropy bands. Here, we establish two properties that motivate these metrics:

- **Frequency-weighted decomposition.** The overall correctness AUROC is a frequency-weighted combination of the four atomic AUROCs (Proposition A.1). It can therefore remain high when the dominant entropy-aligned comparisons are easy, while concealing poor performance on less frequent entropy-misaligned comparisons.
- **Behaviour without an entropy shortcut.** If the predicted correctness score is independent of entropy band conditioned on the true correctness label, all four atomic AUROCs equal the

Table 7: Model-wise AUROC of semantic entropy, SEP, and the correctness probe across correctness–entropy subsets. The average rows are unweighted means across models.

Model	Semantic Entropy						SEP						Correctness Probe					
	CvI	CHvIH	CHvIL	CLvIH	CLvIL	xAUC	CvI	CHvIH	CHvIL	CLvIH	CLvIL	xAUC	CvI	CHvIH	CHvIL	CLvIH	CLvIL	xAUC
Text QA																		
NQ																		
Gemma	0.705	0.626	0.000	1.000	0.601	0.000	0.765	0.697	0.325	0.943	0.704	0.325	0.862	0.852	0.692	0.940	0.832	0.692
Llama	0.768	0.653	0.000	1.000	0.674	0.000	0.749	0.630	0.278	0.924	0.697	0.278	0.828	0.786	0.538	0.919	0.773	0.538
Mistral	0.758	0.673	0.000	1.000	0.649	0.000	0.719	0.609	0.244	0.907	0.657	0.244	0.815	0.778	0.564	0.904	0.768	0.564
Qwen	0.769	0.650	0.000	1.000	0.611	0.000	0.784	0.675	0.273	0.938	0.698	0.273	0.851	0.812	0.575	0.934	0.789	0.575
Average	0.750	0.651	0.000	1.000	0.634	0.000	0.754	0.653	0.280	0.928	0.689	0.280	0.839	0.807	0.592	0.924	0.791	0.592
SciQ																		
Gemma	0.653	0.605	0.000	1.000	0.550	0.000	0.783	0.681	0.482	0.894	0.762	0.482	0.871	0.801	0.684	0.931	0.862	0.684
Llama	0.785	0.616	0.000	1.000	0.687	0.000	0.794	0.649	0.351	0.928	0.744	0.351	0.847	0.757	0.620	0.917	0.832	0.620
Mistral	0.759	0.616	0.000	1.000	0.695	0.000	0.753	0.610	0.349	0.902	0.736	0.349	0.831	0.753	0.616	0.907	0.828	0.616
Qwen	0.716	0.633	0.000	1.000	0.563	0.000	0.776	0.645	0.392	0.906	0.749	0.392	0.857	0.817	0.650	0.930	0.828	0.650
Average	0.728	0.618	0.000	1.000	0.624	0.000	0.777	0.646	0.394	0.908	0.748	0.394	0.852	0.782	0.643	0.921	0.838	0.643
TriviaQA																		
Gemma	0.754	0.700	0.000	1.000	0.604	0.000	0.882	0.802	0.559	0.953	0.848	0.559	0.912	0.842	0.673	0.960	0.890	0.673
Llama	0.881	0.707	0.000	1.000	0.752	0.000	0.875	0.740	0.452	0.946	0.809	0.452	0.898	0.798	0.575	0.951	0.850	0.575
Mistral	0.858	0.723	0.000	1.000	0.724	0.000	0.863	0.748	0.435	0.943	0.793	0.435	0.898	0.828	0.629	0.949	0.852	0.629
Qwen	0.827	0.705	0.000	1.000	0.679	0.000	0.879	0.777	0.488	0.954	0.823	0.488	0.913	0.847	0.644	0.963	0.876	0.644
Average	0.830	0.709	0.000	1.000	0.690	0.000	0.875	0.767	0.484	0.949	0.818	0.484	0.905	0.829	0.630	0.956	0.867	0.630
Image QA																		
AdvQA																		
Gemma	0.646	0.600	0.000	1.000	0.530	0.000	0.737	0.678	0.281	0.917	0.699	0.281	0.777	0.702	0.502	0.874	0.766	0.502
Pixtral	0.744	0.629	0.000	1.000	0.676	0.000	0.742	0.675	0.317	0.918	0.667	0.317	0.761	0.646	0.446	0.878	0.745	0.446
QwenV	0.732	0.620	0.000	1.000	0.702	0.000	0.689	0.621	0.198	0.896	0.616	0.198	0.795	0.740	0.558	0.893	0.783	0.558
Average	0.707	0.616	0.000	1.000	0.636	0.000	0.723	0.658	0.265	0.910	0.661	0.265	0.778	0.696	0.502	0.882	0.765	0.502
OK-VQA																		
Gemma	0.634	0.578	0.000	1.000	0.550	0.000	0.732	0.605	0.290	0.921	0.727	0.290	0.849	0.821	0.675	0.929	0.841	0.675
Pixtral	0.745	0.598	0.000	1.000	0.677	0.000	0.776	0.692	0.339	0.932	0.727	0.339	0.823	0.765	0.560	0.918	0.796	0.560
QwenV	0.743	0.619	0.000	1.000	0.678	0.000	0.741	0.626	0.144	0.966	0.669	0.144	0.830	0.784	0.457	0.950	0.777	0.457
Average	0.707	0.598	0.000	1.000	0.635	0.000	0.750	0.641	0.258	0.940	0.708	0.258	0.834	0.790	0.564	0.932	0.805	0.564
VQAv2																		
Gemma	0.664	0.605	0.000	1.000	0.548	0.000	0.788	0.679	0.311	0.952	0.770	0.311	0.862	0.814	0.597	0.953	0.851	0.597
Pixtral	0.674	0.466	0.000	1.000	0.701	0.000	0.765	0.774	0.560	0.877	0.704	0.560	0.841	0.835	0.699	0.913	0.815	0.699
QwenV	0.763	0.593	0.000	1.000	0.680	0.000	0.754	0.582	0.145	0.955	0.707	0.145	0.833	0.755	0.469	0.940	0.806	0.469
Average	0.700	0.555	0.000	1.000	0.643	0.000	0.769	0.678	0.339	0.928	0.727	0.339	0.845	0.801	0.588	0.935	0.824	0.588

$C_a \vee I_b$ denotes $\mathrm{AUC}(C_a, I_b)$, with correct samples treated as positives. **xAUC** denotes ES-xAUC. **Gemma**, **Llama**, **Mistral**, **Qwen**, **Pixtral**, and **QwenV** denote `Gemma-3-12B`, `Llama-3.1-8B`, `Mistral-7B-v0.3`, `Qwen2.5-7B`, `Pixtral-12B`, and `Qwen2.5-VL-7B`, respectively.

overall AUROC. Consequently, ES-xAUC equals the overall AUROC and ES-δAUC equals zero (Proposition A.2).

Let $s(x) = \hat{Z}(x)$ denote the correctness score assigned to sample x, where a larger score indicates a higher predicted probability of correctness. For a correct subset $D \subseteq C$ and an incorrect subset $A \subseteq I$, define

$$\mathrm{AUC}(D, A) = \frac{1}{|D||A|} \sum_{x^+ \in D} \sum_{x^- \in A} \ell\big(s(x^+), s(x^-)\big), \tag{12}$$

where

$$\ell(u, v) = \mathbf{1}\{u > v\} + \frac{1}{2}\mathbf{1}\{u = v\}. \tag{13}$$

Thus, $\mathrm{AUC}(D, A)$ is the probability that a randomly selected correct sample from D receives a higher score than a randomly selected incorrect sample from A, with ties assigned half credit. We assume that every subset appearing in an AUROC computation is nonempty.

Proposition A.1 (Frequency decomposition of overall AUROC)**.** *Let $\pi_a^I = \frac{|I_a|}{|I|}$ and $\pi_b^C = \frac{|C_b|}{|C|}$ for $a, b \in \{L, H\}$. Then the overall correctness AUROC satisfies*

$$\mathrm{AUC}(C, I) = \sum_{a,b \in \{L,H\}} \pi_a^I \pi_b^C \, \mathrm{AUC}(C_b, I_a). \tag{14}$$

Proof. Because the entropy bands form disjoint partitions,

$$I = I_L \,\dot\cup\, I_H, \qquad C = C_L \,\dot\cup\, C_H. \tag{15}$$

Using Equation 12, the overall AUROC can be decomposed as

$$\begin{aligned}\mathrm{AUC}(C, I) &= \frac{1}{|C||I|} \sum_{x^+ \in C} \sum_{x^- \in I} \ell\big(s(x^+), s(x^-)\big) \\ &= \frac{1}{|C||I|} \sum_{a,b \in \{L,H\}} \sum_{x^+ \in C_b} \sum_{x^- \in I_a} \ell\big(s(x^+), s(x^-)\big) \\ &= \sum_{a,b \in \{L,H\}} \frac{|C_b||I_a|}{|C||I|} \,\mathrm{AUC}(C_b, I_a) \\ &= \sum_{a,b \in \{L,H\}} \pi_a^I \pi_b^C \,\mathrm{AUC}(C_b, I_a). \end{aligned} \tag{16}$$

Therefore, the overall AUROC is a frequency-weighted combination of the four atomic AUROCs. □

Proposition A.2 (Metrics under entropy invariance)**.** *Let $S = \hat{Z}(X)$ denote the predicted correctness score. If*

$$S \perp B \mid Z, \tag{17}$$

then, at the population level,

$$\begin{aligned}\mathrm{AUC}(C_H, I_H) &= \mathrm{AUC}(C_H, I_L) \\ &= \mathrm{AUC}(C_L, I_H) \\ &= \mathrm{AUC}(C_L, I_L) \\ &= \mathrm{AUC}(C, I). \end{aligned} \tag{18}$$

Consequently,

$$\begin{aligned}\text{ES-xAUC} &= \mathrm{AUC}(C, I), \\ \text{ES-}\delta\text{AUC} &= 0, \\ \mathrm{AUC}(C_H, I_L) &= \mathrm{AUC}(C, I). \end{aligned} \tag{19}$$

Proof. The conditional-independence assumption implies that, for every $a, b \in \{L, H\}$,

$$\begin{aligned} S \mid (Z = 0, B = a) &\overset{d}{=} S \mid Z = 0, \\ S \mid (Z = 1, B = b) &\overset{d}{=} S \mid Z = 1, \end{aligned} \tag{20}$$

where $\overset{d}{=}$ denotes equality in distribution. Therefore, the score distribution of incorrect samples does not depend on their entropy band, and the same holds for correct samples.

Let

$$S_a^- \sim S \mid (Z = 0, B = a), \qquad S_b^+ \sim S \mid (Z = 1, B = b) \tag{21}$$

be independent draws. By Equation 20, the joint distribution of (S_b^+, S_a^-) is identical for every $a, b \in \{L, H\}$ and is equal to the joint distribution obtained by drawing from the complete correct and incorrect populations. The pairwise interpretation of AUROC therefore gives

$$\mathrm{AUC}(C_b, I_a) = \mathrm{AUC}(C, I), \qquad \forall\, a, b \in \{L, H\}. \tag{22}$$

Taking the minimum and maximum over the four identical atomic AUROCs yields

$$\text{ES-xAUC} = \mathrm{AUC}(C, I), \qquad \text{ES-}\delta\text{AUC} = 0. \tag{23}$$

The equality $\mathrm{AUC}(C_H, I_L) = \mathrm{AUC}(C, I)$ follows directly from Equation 18. □

In finite samples, these equalities may hold only approximately because of AUROC estimation variability. Moreover, equality of the four atomic AUROCs is a necessary consequence of $S \perp B \mid Z$, but it is not sufficient to establish full conditional independence.

A.3 LLM-AS-A-JUDGE

Because LLMs generate free-form answers, lexically different responses can nevertheless be semantically equivalent. We therefore use Llama-3.3-70B-Instruct Grattafiori et al. (2024) as a fixed external judge to determine whether the model prediction matches a reference answer. The judge receives the question, the reference answers, and only the model's greedy prediction. The stochastic generations are used solely to compute semantic entropy and are not judged.

The placeholder `{question}` contains the stripped dataset question, `{ground_truth}` contains the numbered reference answers, and `{prediction}` contains the model's greedy answer. Reference answers and predictions are lower-cased and stripped; non-alphanumeric ASCII characters are removed and consecutive whitespace is collapsed. The question itself is not normalised. The judge response is lower-cased and parsed by searching first for `yes` and then for `no`. For image-QA datasets, the judge is text-only and does not receive the image. The exact prompts are provided below.

SciQ judge prompt

```
[SYSTEM]
You are an expert science answer evaluator.
IMPORTANT: Evaluate the proposed answer IN THE CONTEXT OF WHAT THE QUESTION ASKS.
The proposed answer is correct if it conveys the same scientific meaning as any of the valid answers.
Focus on the meaning being conveyed, not the surface form -- synonyms, paraphrases, and abbreviations are
      acceptable.
If a numeric answer is present in the ground truth, focus on the value only -- formatting differences such
      as 5.0 vs 5, $100 vs 100, 1,000 vs 1000, or 4 vs four are the same.
A partial or abbreviated answer is acceptable if it unambiguously identifies the correct answer given the
      question context.
There may be multiple valid answers listed -- the proposed answer is correct if it matches ANY one of them.
Respond with exactly one word: yes or no.

[USER]
Question: {question}
Valid answer(s) (match ANY one):
{ground_truth}
Proposed answer: {prediction}

In the context of the question asked, does the proposed answer convey the same scientific meaning as at
      least one of the valid answers? Respond with exactly one word: yes or no.
```

TriviaQA judge prompt

```
[SYSTEM]
You are an expert factual QA evaluator.
IMPORTANT: Evaluate the proposed answer IN THE CONTEXT OF WHAT THE QUESTION ASKS.
The proposed answer is correct if it refers to the same entity or concept as any of the valid answers.
Focus on the meaning and identity being conveyed -- aliases, abbreviations, and common name variations are
      acceptable.
If the answer is numerical, focus on the value being expressed -- formatting differences such as 5.0 vs 5,
      $100 vs 100, 1,000 vs 1000, or 4 vs four are the same.
A partial or abbreviated answer is acceptable if it unambiguously identifies the correct entity given the
      question context.
There may be multiple valid answers listed -- the proposed answer is correct if it matches ANY one of them.
Respond with exactly one word: yes or no.

[USER]
Question: {question}
Valid answer(s) (match ANY one):
{ground_truth}
Proposed answer: {prediction}

In the context of the question asked, does the proposed answer refer to the same entity or concept as at
      least one of the valid answers? Respond with exactly one word: yes or no.
```

Natural Questions judge prompt

```
[SYSTEM]
You are an expert factual QA evaluator.
IMPORTANT: Evaluate the proposed answer IN THE CONTEXT OF WHAT THE QUESTION ASKS.
The proposed answer is correct if it refers to the same entity, date, or concept as any of the valid answers.
Focus on the meaning and identity being conveyed -- aliases, abbreviations, and common name variations are
      acceptable.
For dates, a coarser but consistent answer is acceptable if it does not contradict a valid answer (December
      1972 vs 14 December 1972 are the same; December 1971 is not).
```

```
If the answer is numerical, focus on the value being expressed -- formatting differences such as 5.0 vs 5,
      $100 vs 100, 1,000 vs 1000, or 4 vs four are the same.
A partial or abbreviated answer is acceptable if it unambiguously identifies the correct entity given the
      question context.
There may be multiple valid answers listed -- the proposed answer is correct if it matches ANY one of them.
Respond with exactly one word: yes or no.

[USER]
Question: {question}
Valid answer(s) (match ANY one):
{ground_truth}
Proposed answer: {prediction}

In the context of the question asked, does the proposed answer refer to the same entity, date, or concept as
      at least one of the valid answers? Respond with exactly one word: yes or no.
```

AdVQA, OK-VQA, and VQAv2 judge prompt

```
[SYSTEM]
You are an expert visual question answering (VQA) evaluator.
IMPORTANT: Evaluate the proposed answer IN THE CONTEXT OF WHAT THE QUESTION ASKS.
The proposed answer is correct if it conveys the same meaning as any of the valid answers. These are short,
      free-form answers provided by human annotators viewing the image, and are often a single word or short
      phrase.
Focus on the meaning being conveyed, not the surface form -- synonyms, paraphrases, singular/plural
      variants, and minor wording differences are acceptable. Numerical answers should be compared by value
      (5.0 vs 5 and 4 vs four are the same).
There may be multiple valid answers listed -- the proposed answer is correct if it matches ANY one of them.
Respond with exactly one word: yes or no.

[USER]
Question about an image: {question}
Valid answer(s) (match ANY one):
{ground_truth}
Proposed answer: {prediction}

In the context of the question asked, does the proposed answer convey the same meaning as at least one of
      the valid answers? Respond with exactly one word: yes or no.
```

A.4 LLM PROBES

Linear probes on model internals are among the strongest post-hoc correctness predictors for LLMs. They leave the underlying LLM or VLM frozen and train only a lightweight classifier using features extracted during the greedy forward pass. We evaluate the following probe families, which differ in either their training target or the internal representation provided to the classifier:

- **Semantic Entropy Probe (SEP).** SEP Han et al. (2024) takes the model hidden states as input but is trained to predict the binary semantic-entropy band $B \in \{L, H\}$ rather than correctness. For correctness detection, its score is oriented such that a larger value indicates greater probability of belonging to the low-entropy band. SEP therefore provides a direct representation-level baseline for measuring the extent to which entropy is encoded in the hidden states.
- **Correctness Probe.** Following Orgad et al. Orgad et al. (2025), the correctness probe takes hidden-state representations as input and is trained directly using the correctness label Z. We extract representations from middle layers, spanning approximately 35%–69% of model depth, and late layers, spanning the final approximately 31%. Representations are averaged within each depth range, and the two resulting vectors are concatenated before being provided to the classifier.
- **Attention Sink (SINK).** The SINK probe uses attention-sink statistics as its input representation Binkowski et al. (2026). For each attention head, it extracts the largest attention-sink scores, capturing the extent to which attention concentrates on a small set of tokens. These features are concatenated across heads and used to predict correctness.
- **Laplacian Eigenvalues (LapEigVals).** This probe represents each attention head as an attention graph and uses the leading eigenvalues of its graph Laplacian as features. These eigenvalues summarize the global spectral structure of the attention pattern and have been proposed as internal signals for hallucination detection Binkowski et al. (2025).

- **Attention Log-Determinant (AttnLogDet).** AttnLogDet summarizes each attention head using a log-determinant statistic derived from its attention matrix. The resulting values across heads and layers form the input to the correctness classifier Sriramanan et al. (2024).
- **Attention Eigenvalues (AttnEigVals).** AttnEigVals uses the leading eigenvalues of each attention matrix directly, rather than those of its graph Laplacian. The eigenvalue features from the selected heads and layers are concatenated and used to train a correctness probe Binkowski et al. (2025).

Each feature family is evaluated separately and is provided to a class-balanced logistic-regression classifier. All probes except SEP are trained using correctness labels; SEP is trained using entropy-band labels. The underlying LLM or VLM remains frozen in every case, ensuring that the comparison isolates the information captured by the different internal representations.

A.5 Joint Optimisation

Joint optimisation trains a single correctness probe to perform consistently across entropy bands. Unlike routed learning, it does not use separate band-specific experts. Instead, the entropy bands are used only to structure the training loss; the resulting probe applies the same scoring function to every test sample and does not require entropy-band information at inference time.

Let

$$p_\theta(X) = \sigma\big(\mathbf{w}^\top h_{\mathcal{M}}(X) + \beta\big), \qquad \theta = (\mathbf{w}, \beta), \tag{24}$$

denote the correctness probability predicted from the frozen model representation, where $Z = 1$ denotes a correct prediction and $Z = 0$ an incorrect prediction. We partition the training set into

$$\mathcal{G} = \{C_H, C_L, I_H, I_L\}. \tag{25}$$

For each group $g \in \mathcal{G}$, its mean binary cross-entropy loss is

$$\mathcal{L}_g(\theta) = -\frac{1}{|\mathcal{D}_g|} \sum_{i \in \mathcal{D}_g} \Big[Z_i \log p_\theta(X_i) + (1 - Z_i) \log(1 - p_\theta(X_i)) \Big]. \tag{26}$$

Under ordinary empirical-risk minimisation, these group losses are implicitly weighted by their frequencies:

$$\mathcal{L}_{\mathrm{ERM}}(\theta) = \sum_{g \in \mathcal{G}} \rho_g \, \mathcal{L}_g(\theta), \qquad \rho_g = \frac{|\mathcal{D}_g|}{|\mathcal{D}|}. \tag{27}$$

Consequently, the dominant entropy-aligned groups, C_L and I_H, can dominate the learned decision rule, while the less frequent entropy-misaligned groups, C_H and I_L, contribute comparatively little. We consider two alternatives that remove this frequency dependence.

`BCE-Equal`. The first approach assigns equal importance to all four groups:

$$\theta^*_{\mathrm{equal}} = \arg\min_\theta \frac{1}{4} \sum_{g \in \mathcal{G}} \mathcal{L}_g(\theta). \tag{28}$$

Because each $\mathcal{L}_g$ is an average within its group, every correctness–entropy subgroup contributes equally irrespective of its sample count. This prevents the naturally larger C_L and I_H groups from overwhelming the entropy-misaligned groups during training.

`BCE-DRO`. Equal weighting does not account for differences in group difficulty. We therefore also use Group DRO, which adaptively emphasizes groups with larger training loss:

$$\theta^*_{\mathrm{DRO}} = \arg\min_\theta \max_{\mathbf{q} \in \Delta_4} \sum_{g \in \mathcal{G}} q_g \, \mathcal{L}_g(\theta), \tag{29}$$

where

$$\Delta_4 = \left\{ \mathbf{q} \in \mathbb{R}^4_+ : \sum_{g \in \mathcal{G}} q_g = 1 \right\}. \tag{30}$$

The inner maximisation assigns greater weight to groups on which the current probe incurs larger BCE. We implement this using exponentiated-gradient updates. Starting from $q_{g,0} = 1/4$, the group weights are updated as

$$q_{g,t+1} = \frac{q_{g,t} \exp(\eta_q \mathcal{L}_g(\theta_t))}{\sum_{g' \in \mathcal{G}} q_{g',t} \exp(\eta_q \mathcal{L}_{g'}(\theta_t))}, \tag{31}$$

followed by an update of the shared logistic-regression parameters:

$$\theta_{t+1} = \theta_t - \eta_\theta \nabla_\theta \left[\sum_{g \in \mathcal{G}} q_{g,t+1} \mathcal{L}_g(\theta_t) \right]. \tag{32}$$

Thus, the logistic regressor is not independently refitted after every weight update. Rather, the group weights and the parameters of the same probe are updated iteratively during training. Groups that remain difficult accumulate larger weights, directing subsequent parameter updates toward their losses. Importantly, Group DRO optimizes the worst-group BCE rather than directly optimizing AUROC, ES-xAUC, or ES-δAUC.

Comparison of the joint objectives. Table 8 reports the complete model–dataset-level results for `BCE-Equal` and `BCE-DRO`. Relative to equal weighting, Group DRO improves ES-xAUC and reduces ES-δAUC in 20 of the 21 model–dataset combinations. The only exception is `Qwen2.5-VL-7B` on AdvQA, for which ES-xAUC decreases from 0.670 to 0.619 and ES-δAUC increases from 0.150 to 0.228.

Averaged over the text-QA models and datasets, moving from `BCE-Equal` to `BCE-DRO` increases ES-xAUC from 69.2% to 72.0% and reduces ES-δAUC from 20.9% to 15.1%. For image QA, it increases ES-xAUC from 63.5% to 67.0% and reduces ES-δAUC from 23.5% to 15.4%. This increased robustness is accompanied by a modest reduction in pooled correctness AUROC: CvI changes from 85.1% to 83.5% for text QA and from 80.2% to 78.1% for image QA. This reflects the intended trade-off of worst-group optimisation: performance is redistributed toward the difficult entropy-misaligned groups rather than concentrated on the dominant groups.

Across every row of Table 8, $\mathrm{AUC}(C_H, I_L)$ remains the minimum atomic AUROC and therefore determines ES-xAUC. Nevertheless, Group DRO substantially raises this bottleneck relative to the standard Correctness Probe. Averaged across datasets, ES-xAUC increases from 62.2% to 72.0% for text QA and from 55.1% to 67.0% for image QA, while ES-δAUC decreases from 31.2% to 15.1% and from 36.5% to 15.4%, respectively.

We therefore use `BCE-DRO` as the proposed *Joint Learning* method in the main experiments. Table 14 compares it with the prior probe-based approaches and Routed Learning at the dataset-averaged level. Table 15 reports the corresponding atomic AUROCs for every model–dataset combination, providing a direct model-wise comparison between Joint Learning and Routed Learning.

A.6 Routed learning

Joint learning asks a single correctness probe to remain stable across entropy bands. Routed learning instead prevents the correctness predictor from comparing high- and low-entropy samples during training. It comprises three components: (i) a binary router that assigns each input to an entropy band, (ii) separate correctness probes trained within the high- and low-entropy bands, and (iii) bandwise normalization that makes the two experts' scores comparable before pooling.

Specifically, the high-entropy expert is trained on $C_H \cup I_H$, while the low-entropy expert is trained on $C_L \cup I_L$. The entropy band is therefore constant within each expert's training pool and cannot itself serve as a varying feature for predicting correctness.

A.6.1 Theoretical foundation

Let

$$s_b(X) = \mathbf{w}_b^\top h_{\mathcal{M}}(X) + \beta_b, \qquad b \in \{H, L\}, \tag{33}$$

denote the pre-sigmoid correctness logit produced by the expert assigned to entropy band b. Let

$$\pi_b = \Pr(Z = 1 \mid B = b) \tag{34}$$

Table 8: Model–dataset-level atomic AUROC results for BCE-Equal (equal-weighted BCE across the four atomic cells) and BCE-DRO (Group-DRO-weighted BCE across the four atomic cells) across text and image QA. CvI is the pooled correct-vs-incorrect AUROC. Band AUC reports the four atomic comparisons. ES-xAUC is their minimum, while ES-δAUC is their maximum-to-minimum gap.

Dataset	Model	BCE-Equal							BCE-DRO						
		CvI	Band AUC				ES-xAUC	ES-δAUC	CvI	Band AUC				ES-xAUC	ES-δAUC
		↑	C_H I_H	C_H I_L	C_L I_H	C_L I_L	↑	↓	↑	C_H I_H	C_H I_L	C_L I_H	C_L I_L	↑	↓
Text QA															
NQ	Gemma	0.846	0.855	0.751	0.906	0.819	0.751	0.155	0.846	0.843	0.776	0.888	0.828	0.776	0.111
	Llama	0.814	0.789	0.653	0.870	0.770	0.653	0.218	0.791	0.781	0.709	0.820	0.756	0.709	0.111
	Mistral	0.801	0.777	0.671	0.853	0.768	0.671	0.182	0.786	0.771	0.707	0.819	0.762	0.707	0.112
	Qwen	0.830	0.810	0.691	0.880	0.784	0.691	0.188	0.808	0.799	0.733	0.838	0.777	0.733	0.104
SciQ	Gemma	0.861	0.802	0.720	0.909	0.854	0.720	0.190	0.857	0.792	0.723	0.901	0.851	0.723	0.178
	Llama	0.820	0.747	0.683	0.865	0.817	0.683	0.182	0.802	0.740	0.703	0.834	0.806	0.703	0.131
	Mistral	0.819	0.756	0.682	0.871	0.822	0.682	0.189	0.794	0.750	0.720	0.822	0.804	0.720	0.102
	Qwen	0.849	0.829	0.709	0.907	0.819	0.709	0.197	0.838	0.827	0.734	0.884	0.808	0.734	0.150
TriviaQA	Gemma	0.906	0.834	0.698	0.952	0.885	0.698	0.254	0.897	0.816	0.717	0.936	0.880	0.717	0.219
	Llama	0.880	0.789	0.631	0.926	0.836	0.631	0.295	0.855	0.777	0.654	0.895	0.812	0.654	0.241
	Mistral	0.885	0.828	0.700	0.924	0.846	0.700	0.224	0.863	0.813	0.727	0.895	0.831	0.727	0.168
	Qwen	0.901	0.848	0.709	0.943	0.867	0.709	0.234	0.881	0.828	0.736	0.915	0.854	0.736	0.179
Image QA															
AdvQA	Gemma	0.751	0.675	0.511	0.838	0.741	0.511	0.327	0.752	0.679	0.560	0.819	0.746	0.560	0.258
	Pixtral	0.728	0.605	0.491	0.818	0.726	0.491	0.327	0.690	0.570	0.526	0.747	0.707	0.526	0.222
	QwenV	0.773	0.730	0.670	0.820	0.773	0.670	0.150	0.778	0.739	0.619	0.848	0.763	0.619	0.228
OK-VQA	Gemma	0.844	0.822	0.736	0.897	0.837	0.736	0.161	0.825	0.808	0.782	0.844	0.824	0.782	0.062
	Pixtral	0.805	0.755	0.621	0.877	0.782	0.621	0.256	0.773	0.738	0.671	0.816	0.757	0.671	0.146
	QwenV	0.808	0.773	0.623	0.880	0.775	0.623	0.257	0.803	0.776	0.639	0.865	0.762	0.639	0.226
VQAv2	Gemma	0.859	0.813	0.679	0.925	0.849	0.679	0.247	0.822	0.780	0.758	0.838	0.828	0.758	0.080
	Pixtral	0.839	0.833	0.709	0.906	0.814	0.709	0.198	0.811	0.803	0.731	0.856	0.793	0.731	0.125
	QwenV	0.811	0.757	0.671	0.863	0.804	0.671	0.192	0.773	0.752	0.745	0.785	0.781	0.745	0.040

CvI is the pooled correct-vs-incorrect AUROC. Within each Band-AUC block, the columns report $\mathrm{AUC}(C_H, I_H)$, $\mathrm{AUC}(C_H, I_L)$, $\mathrm{AUC}(C_L, I_H)$, and $\mathrm{AUC}(C_L, I_L)$, respectively. Light-red shading highlights the entropy-misaligned $\mathrm{AUC}(C_H, I_L)$ comparison and the two summary metrics. **Gemma**, **Llama**, **Mistral**, **Qwen**, **Pixtral**, and **QwenV** denote `Gemma-3-12B`, `Llama-3.1-8B`, `Mistral-7B-v0.3`, `Qwen2.5-7B`, `Pixtral-12B`, and `Qwen2.5-VL-7B`, respectively.

denote the correctness rate within that band. Under the natural distribution, correct predictions occur predominantly in the low-entropy band, whereas incorrect predictions occur predominantly in the high-entropy band. Consequently, $\pi_L > \pi_H$.

Proposition A.3 (Routing absorbs the entropy shortcut into the intercept)**.** *Suppose that the hidden representation contains a correctness-evidence score whose class-conditional likelihood ratio is invariant across entropy bands:*

$$r(h) = \log \frac{p(h \mid Z = 1, B = b)}{p(h \mid Z = 0, B = b)} = \mathbf{w}_c^\top h + \beta_c, \qquad b \in \{H, L\}. \tag{35}$$

Then the Bayes-optimal correctness logit within band b is

$$s_b^*(h) = \mathbf{w}_c^\top h + \beta_c + \log \frac{\pi_b}{1 - \pi_b}. \tag{36}$$

Thus, both band-specific experts share the correctness direction $\mathbf{w}_c$, while the entropy–correctness association is represented by the band-specific intercept

$$\beta_b = \beta_c + \mathrm{logit}(\pi_b). \tag{37}$$

In particular,

$$\beta_L - \beta_H = \log \frac{\pi_L(1 - \pi_H)}{\pi_H(1 - \pi_L)} > 0. \tag{38}$$

Proof. By Bayes' rule,

$$\begin{aligned} \mathrm{logit}(\Pr(Z = 1 \mid h, B = b)) &= \log \frac{p(h \mid Z = 1, B = b)}{p(h \mid Z = 0, B = b)} + \log \frac{\Pr(Z = 1 \mid B = b)}{\Pr(Z = 0 \mid B = b)} \\ &= r(h) + \log \frac{\pi_b}{1 - \pi_b}. \end{aligned} \tag{39}$$

Substituting Equation 35 gives Equation 36. Because $B = b$ is constant within each expert's training pool, the band-dependent prior cannot define a varying decision direction and is instead absorbed by the intercept. □

The proposition also explains why a dedicated expert need not improve within-band AUROC. For $h_b^+ \sim C_b$ and $h_b^- \sim I_b$,

$$\begin{aligned}\Pr\big(s_b^*(h_b^+) > s_b^*(h_b^-)\big) &= \Pr\big(r(h_b^+) + \mathrm{logit}(\pi_b) > r(h_b^-) + \mathrm{logit}(\pi_b)\big) \\ &= \Pr\big(r(h_b^+) > r(h_b^-)\big)\,. \end{aligned} \tag{40}$$

The additive intercept affects neither rankings nor ties within a band. It therefore does not change $\mathrm{AUC}(C_H, I_H)$ or $\mathrm{AUC}(C_L, I_L)$.

The intercept does, however, affect comparisons across experts. For the entropy-aligned comparison,

$$\begin{aligned}s_L^*(h_{C_L}) - s_H^*(h_{I_H}) = r(h_{C_L}) - r(h_{I_H}) \\ + \mathrm{logit}(\pi_L) - \mathrm{logit}(\pi_H),\end{aligned} \tag{41}$$

so the natural entropy–correctness association increases $\mathrm{AUC}(C_L, I_H)$. Conversely,

$$\begin{aligned}s_H^*(h_{C_H}) - s_L^*(h_{I_L}) = r(h_{C_H}) - r(h_{I_L}) \\ - \left[\mathrm{logit}(\pi_L) - \mathrm{logit}(\pi_H)\right],\end{aligned} \tag{42}$$

which penalizes the entropy-misaligned $\mathrm{AUC}(C_H, I_L)$. Directly pooling the raw logits therefore retains the entropy shortcut through their band-dependent offsets.

A.6.2 Bandwise normalization and score pooling

To place the two experts on a common score scale, we estimate the mean and standard deviation of each expert's logits on its corresponding training pool:

$$\begin{aligned}\mu_b &= \mathbb{E}_{\mathrm{train}}[s_b(X) \mid B = b]\,, \\ \sigma_b &= \sqrt{\mathrm{Var}_{\mathrm{train}}[s_b(X) \mid B = b]}.\end{aligned} \tag{43}$$

The normalized expert score is

$$\widetilde{s}_b(X) = \frac{s_b(X) - \mu_b}{\sigma_b}. \tag{44}$$

Under Proposition A.3, substituting Equation 36 gives

$$\widetilde{s}_b^*(X) = \frac{r(h_{\mathcal{M}}(X)) - \mathbb{E}_{\mathrm{train}}[r(h_{\mathcal{M}}(X)) \mid B = b]}{\sqrt{\mathrm{Var}_{\mathrm{train}}[r(h_{\mathcal{M}}(X)) \mid B = b]}}. \tag{45}$$

The additive term $\mathrm{logit}(\pi_b)$ therefore cancels exactly. More generally, normalization removes the fitted band-specific location and rescales the logits using statistics estimated exclusively from the training split.

Let $\widehat{B}(X)$ denote the entropy band predicted by the router. The final routed correctness score is

$$s_{\mathrm{route}}(X) = \widetilde{s}_{\widehat{B}(X)}(X). \tag{46}$$

At test time, each example is scored by only the expert selected by the router and normalized using that expert's stored training-set statistics. No test-set statistics are used.

A.6.3 Training and performance of the entropy-band router

The router is a binary linear classifier trained to predict the entropy band $B \in \{L, H\}$. The target bands are obtained by thresholding discrete semantic entropy as described in Appendix A.8.1. For each input, the router receives the sampling-cloud representation constructed from the hidden states of the $n = 10$ stochastic generations. The router is fitted only on the natural training split and is then applied unchanged to held-out examples.

Table 9 reports model–dataset-level router AUROC. The mean AUROC is 0.954 for text QA and 0.921 for image QA. Dataset-level means range from 0.944 to 0.971 for text QA and from 0.900 to 0.932 for image QA. Thus, the sampled representation reliably identifies the entropy band across models and datasets, although routing remains an estimated rather than oracle operation.

Table 9: Model–dataset-level entropy-band router AUROC using the sampling-cloud representation constructed from $n = 10$ high-temperature generations. The router is trained on the `Natural` training split. The final column averages across datasets, and the final row of each panel averages across models.

Text QA				
Model	**NQ**	**SciQ**	**TriviaQA**	**Mean**
Gemma-3-12B	0.957	0.954	0.979	0.963
Llama-3.1-8B	0.939	0.950	0.957	0.949
Mistral-7B-v0.3	0.929	0.931	0.961	0.940
Qwen2.5-7B	0.949	0.953	0.985	0.962
Mean	**0.944**	**0.947**	**0.971**	**0.954**
Image QA				
Model	**VQAv2**	**OK-VQA**	**AdvQA**	**Mean**
Gemma-3-12B	0.914	0.908	0.897	0.906
Pixtral-12B	0.948	0.941	0.933	0.941
Qwen2.5-VL-7B	0.931	0.947	0.869	0.916
Mean	**0.931**	**0.932**	**0.900**	**0.921**

Table 10: Own-band correctness AUROC of each entropy-band expert (Exp.) and the standard generalist Correctness Probe (Gen.). Each cell reports *Exp./Gen.*. Band L evaluates C_L against I_L, while Band H evaluates C_H against I_H. Dataset means are shown in the final two columns, and the final row of each panel averages across models.

Text QA								
Model	**NQ**		**SciQ**		**TriviaQA**		**Mean**	
	L	**H**	**L**	**H**	**L**	**H**	**L**	**H**
Gemma-3-12B	0.834/0.832	0.833/0.852	0.857/0.862	0.779/0.801	0.886/0.890	0.819/0.842	0.859/0.861	0.810/0.832
Llama-3.1-8B	0.759/0.773	0.795/0.786	0.808/0.832	0.751/0.757	0.820/0.850	0.793/0.798	0.796/0.818	0.780/0.781
Mistral-7B-v0.3	0.749/0.768	0.787/0.778	0.815/0.828	0.746/0.753	0.843/0.852	0.822/0.828	0.803/0.816	0.785/0.786
Qwen2.5-7B	0.786/0.789	0.802/0.812	0.816/0.828	0.790/0.817	0.875/0.876	0.835/0.847	0.825/0.831	0.809/0.825
Mean	**0.782/0.791**	**0.804/0.807**	**0.824/0.837**	**0.766/0.782**	**0.856/0.867**	**0.817/0.829**	**0.821/0.832**	**0.796/0.806**
Image QA								
Model	**VQAv2**		**OK-VQA**		**AdvQA**		**Mean**	
	L	**H**	**L**	**H**	**L**	**H**	**L**	**H**
Gemma-3-12B	0.848/0.851	0.808/0.814	0.836/0.841	0.803/0.821	0.760/0.766	0.677/0.702	0.815/0.819	0.763/0.779
Pixtral-12B	0.810/0.815	0.830/0.835	0.790/0.796	0.749/0.765	0.737/0.745	0.595/0.646	0.779/0.786	0.725/0.749
Qwen2.5-VL-7B	0.806/0.806	0.757/0.755	0.766/0.777	0.777/0.784	0.779/0.783	0.718/0.740	0.784/0.789	0.751/0.760
Mean	**0.821/0.824**	**0.799/0.801**	**0.797/0.805**	**0.776/0.790**	**0.759/0.765**	**0.663/0.696**	**0.793/0.798**	**0.746/0.762**

A.6.4 WITHIN-BAND PERFORMANCE OF THE CORRECTNESS EXPERTS

We next examine whether routed learning succeeds because the dedicated experts are individually better correctness predictors within their respective entropy bands. The high-entropy expert is evaluated only on C_H and I_H, while the low-entropy expert is evaluated only on C_L and I_L. Each is compared against the standard Correctness Probe trained on all samples, using the same held-out rows and without refitting either model.

Table 10 reports each entry as *expert/generalist*. The generalist has the higher AUROC in 37 of the 42 displayed comparisons; the band expert is higher in four comparisons, with one tie. Averaged across text-QA pairs, the expert-minus-generalist gap is -0.011 in the low-entropy band and -0.010 in the high-entropy band. For image QA, the corresponding gaps are -0.005 and -0.016. The largest reduction is observed for the Pixtral-12B high-entropy expert on AdvQA, with a gap of -0.051.

These results show that routed learning does not succeed because either expert is intrinsically sharper than the generalist within its own band. Indeed, each expert is trained on fewer examples and is consequently slightly weaker on average. The routed improvement instead arises from preventing entropy-band differences from entering the correctness direction and from aligning the two experts' score distributions before pooling.

A.6.5 CORRECTNESS-DETECTION PERFORMANCE

The model-averaged results are reported in Table 14, while complete model–dataset results are reported in Table 15. Routed learning improves all three model-averaged entropy-shortcut metrics over joint learning. On text QA, $\mathrm{AUC}(C_H, I_L)$ increases from 0.720 to 0.817, ES-xAUC increases from 0.720 to 0.736, and ES-δAUC decreases from 0.151 to 0.082. These correspond to improvements of 9.7, 1.6, and 6.9 percentage points, respectively. On image QA, $\mathrm{AUC}(C_H, I_L)$ increases from 0.670 to 0.742, ES-xAUC increases from 0.670 to 0.715, and ES-δAUC decreases from 0.154 to 0.055, corresponding to improvements of 7.2, 4.5, and 9.9 percentage points.

The model–dataset-level results show that these improvements are broadly consistent rather than being driven by a small number of averaged cases. Routed learning improves the entropy-misaligned $\mathrm{AUC}(C_H, I_L)$ in all 21 model–dataset pairs and reduces ES-δAUC in 20 of the 21 pairs. ES-xAUC improves in 14 of the 21 individual pairs and, after averaging across models, improves on all six datasets. Together with the own-band results above, this demonstrates that the principal advantage of routed learning is improved comparability across entropy bands rather than stronger within-band discrimination.

A.7 MATCHED ENTROPY TRAINING

The `Natural` training split contains a strong association between correctness and semantic entropy: correct predictions occur predominantly in the low-entropy band, whereas incorrect predictions occur predominantly in the high-entropy band. We construct an entropy-`Matched` training split to remove this association while retaining as many examples as possible. This provides a controlled intervention: the probe architecture, representation, regularization, and evaluation set remain unchanged, while the availability of entropy as a shortcut is progressively reduced.

A.7.1 CONSTRUCTION OF THE MATCHED SPLIT

Exact entropy strata. Let E_i denote the raw semantic entropy of example i, and let $B_i \in \{L, H\}$ denote its entropy band. Within each band, correct and incorrect examples are grouped by their exact entropy value:

$$\begin{aligned} \mathcal{S}^C_{b,e} &= \{i : Z_i = 1,\ B_i = b,\ E_i = e\}, \\ \mathcal{S}^I_{b,e} &= \{i : Z_i = 0,\ B_i = b,\ E_i = e\}, \qquad b \in \{L, H\}. \end{aligned} \tag{47}$$

No entropy binning or decimal rounding is applied. The numerically equivalent values -0.0 and 0.0 are mapped to the same key.

Constrained sampling. Exact 1:1 matching can discard unnecessarily many examples when the correctness classes are imbalanced. We therefore consider coprime correct-to-incorrect ratios p:q, where $1 \le p, q \le 12$. For entropy stratum (b, e), we retain $p\, t_{b,e}$ correct examples and $q\, t_{b,e}$ incorrect examples, where

$$t_{b,e}(p, q) = \min \left\{ \left\lfloor \frac{|\mathcal{S}^C_{b,e}|}{p} \right\rfloor, \left\lfloor \frac{|\mathcal{S}^I_{b,e}|}{q} \right\rfloor \right\}. \tag{48}$$

A stratum contributes no examples when $t_{b,e} = 0$.

We use the same candidate ratio p:q in both entropy bands. For each candidate, we calculate the retained correct and incorrect counts N^C_b and N^I_b and measure their usable two-class sample size by $n_{\mathrm{eff},b} = (1/N^C_b + 1/N^I_b)^{-1}$. Among ratios whose effective sample size is no smaller than that obtained by 1:1 matching in either band, we select the ratio maximizing $n_{\mathrm{eff},L} + n_{\mathrm{eff},H}$. The 1:1 candidate is always feasible, ensuring that a valid solution exists.

Using a common ratio across the two bands is important. For $T = \sum_{b,e} t_{b,e}(p^\star, q^\star)$, every retained stratum satisfies

$$\widehat{P}(B = b, E = e \mid C) = \frac{p^\star t_{b,e}}{p^\star T} = \frac{t_{b,e}}{T} = \frac{q^\star t_{b,e}}{q^\star T} = \widehat{P}(B = b, E = e \mid I). \tag{49}$$

Thus, entropy and entropy-band membership cannot discriminate between the correct and incorrect classes in the matched training distribution. Choosing different class ratios in the two bands would

Table 11: Detailed dataset statistics and atomic correctness–entropy subsets across text and image QA benchmarks, under the entropy-MATCHED training split. Each entry reports the sample count, followed by the percentage in parentheses. Percentages for C and I are computed relative to N; percentages for C_L and C_H are computed relative to C; and percentages for I_L and I_H are computed relative to I.

Text QA (MATCHED)												
	Natural Questions				**SciQ**				**TriviaQA**			
Statistic	**Gemma**	**Llama**	**Mistral**	**Qwen**	**Gemma**	**Llama**	**Mistral**	**Qwen**	**Gemma**	**Llama**	**Mistral**	**Qwen**
N	9,722	8,850	9,186	8,276	4,979	5,148	5,883	5,716	5,290	4,570	4,890	5,916
C	4,861 (50.0%)	4,425 (50.0%)	4,593 (50.0%)	4,138 (50.0%)	3,491 (70.1%)	3,572 (69.4%)	3,922 (66.7%)	4,287 (75.0%)	2,645 (50.0%)	2,285 (50.0%)	2,445 (50.0%)	2,958 (50.0%)
I	4,861 (50.0%)	4,425 (50.0%)	4,593 (50.0%)	4,138 (50.0%)	1,488 (29.9%)	1,576 (30.6%)	1,961 (33.3%)	1,429 (25.0%)	2,645 (50.0%)	2,285 (50.0%)	2,445 (50.0%)	2,958 (50.0%)
C_L	4,232 (87.1%)	1,383 (31.3%)	1,639 (35.7%)	2,618 (63.3%)	2,660 (76.2%)	1,293 (36.2%)	1,474 (37.6%)	2,265 (52.8%)	2,052 (77.6%)	618 (27.0%)	1,123 (45.9%)	2,168 (73.3%)
C_H	629 (12.9%)	3,042 (68.7%)	2,954 (64.3%)	1,520 (36.7%)	831 (23.8%)	2,279 (63.8%)	2,448 (62.4%)	2,022 (47.2%)	593 (22.4%)	1,667 (73.0%)	1,322 (54.1%)	790 (26.7%)
I_L	4,232 (87.1%)	1,383 (31.3%)	1,639 (35.7%)	2,618 (63.3%)	1,132 (76.1%)	562 (35.7%)	737 (37.6%)	755 (52.8%)	2,052 (77.6%)	618 (27.0%)	1,123 (45.9%)	2,168 (73.3%)
I_H	629 (12.9%)	3,042 (68.7%)	2,954 (64.3%)	1,520 (36.7%)	356 (23.9%)	1,014 (64.3%)	1,224 (62.4%)	674 (47.2%)	593 (22.4%)	1,667 (73.0%)	1,322 (54.1%)	790 (26.7%)
Image QA (MATCHED)												
	AdvQA				**OK-VQA**				**VQAv2**			
Statistic	**Gemma**	**Pixtral**	**QwenV**		**Gemma**	**Pixtral**	**QwenV**		**Gemma**	**Pixtral**	**QwenV**	
N	2,054	1,870	1,890		9,008	7,576	7,804		8,888	9,784	7,326	
C	1,027 (50.0%)	935 (50.0%)	945 (50.0%)		4,504 (50.0%)	3,788 (50.0%)	3,902 (50.0%)		4,444 (50.0%)	5,610 (57.3%)	4,034 (55.1%)	
I	1,027 (50.0%)	935 (50.0%)	945 (50.0%)		4,504 (50.0%)	3,788 (50.0%)	3,902 (50.0%)		4,444 (50.0%)	4,174 (42.7%)	3,292 (44.9%)	
C_L	834 (81.2%)	631 (67.5%)	440 (46.6%)		3,077 (68.3%)	2,023 (53.4%)	774 (19.8%)		2,973 (66.9%)	2,538 (45.2%)	1,136 (28.2%)	
C_H	193 (18.8%)	304 (32.5%)	505 (53.4%)		1,427 (31.7%)	1,765 (46.6%)	3,128 (80.2%)		1,471 (33.1%)	3,072 (54.8%)	2,898 (71.8%)	
I_L	834 (81.2%)	631 (67.5%)	440 (46.6%)		3,077 (68.3%)	2,023 (53.4%)	774 (19.8%)		2,973 (66.9%)	1,882 (45.1%)	922 (28.0%)	
I_H	193 (18.8%)	304 (32.5%)	505 (53.4%)		1,427 (31.7%)	1,765 (46.6%)	3,128 (80.2%)		1,471 (33.1%)	2,292 (54.9%)	2,370 (72.0%)	

C_L and C_H denote correct low- and high-entropy predictions, respectively; I_L and I_H denote incorrect low- and high-entropy predictions. The shaded rows, C_H and I_L, are the entropy-misaligned subsets. **Gemma**, **Llama**, **Mistral**, **Qwen**, **Pixtral**, and **QwenV** denote `Gemma-3-12B`, `Llama-3.1-8B`, `Mistral-7B-v0.3`, `Qwen2.5-7B`, `Pixtral-12B`, and `Qwen2.5-VL-7B`, respectively.

not provide this guarantee: band membership could remain predictive of correctness even if the entropy distributions were matched separately within each band.

Matching is performed only on the training split. All probes trained on `Natural` and `Matched` data are evaluated on the same unchanged natural test split. Consequently, changes in performance reflect changes in the learned probe rather than changes in the evaluation distribution.

A.7.2 MATCHED-SPLIT STATISTICS

Table 11 reports the retained sample counts and correctness–entropy composition for every model–dataset pair. The proportions of low- and high-entropy examples are closely aligned across the correct and incorrect classes, in contrast to the substantial imbalance in the natural split. The resulting dataset need not contain equal numbers of correct and incorrect examples: matching concerns the conditional entropy distributions, not the marginal correctness rate. Allowing a non-1:1 ratio preserves considerably more training data when one correctness class is naturally more frequent.

The intervention nevertheless has a nontrivial sample cost. Across model–dataset pairs, the matched splits retain approximately 30.5%–68.5% of the corresponding natural data. Aggregated across pairs, they retain 45.0% of the text-QA data and 58.4% of the image-QA data. This data reduction motivates the later methods that mitigate the shortcut while training directly on the complete natural split.

A.7.3 EFFECT ON ENTROPY-SUBSET PERFORMANCE

Table 12 reports model–dataset-level results for correctness probes trained on the `Natural` and `Matched` splits. The same frozen representations and probe configuration are used in both conditions.

Table 12: Model- and dataset-level AUROC results corresponding to Table 3. Results are reported for natural and entropy-matched training. ES-xAUC is the minimum AUROC across the four atomic comparisons, while ES-δAUC is their maximum-to-minimum gap.

Dataset	Model	Trained on: Natural							Trained on: Matched						
		CvI	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC	ES-δAUC	CvI	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC	ES-δAUC
Text QA															
SciQ	Gemma-3-12B	0.871	0.801	0.684	0.931	0.862	0.684	0.247	0.862	0.787	0.697	0.913	0.856	0.697	0.216
	Llama-3.1-8B	0.847	0.757	0.620	0.917	0.832	0.620	0.297	0.790	0.740	0.713	0.815	0.793	0.713	0.101
	Mistral-7B-v0.3	0.831	0.753	0.616	0.907	0.828	0.616	0.291	0.788	0.743	0.716	0.816	0.799	0.716	0.100
	Qwen2.5-7B	0.857	0.817	0.650	0.930	0.828	0.650	0.280	0.843	0.834	0.719	0.897	0.809	0.719	0.178
	Average	0.852	0.782	0.643	0.921	0.838	0.643	0.279	0.821	0.776	0.711	0.860	0.815	0.711	0.149
TriviaQA	Gemma-3-12B	0.912	0.842	0.673	0.960	0.890	0.673	0.287	0.898	0.822	0.705	0.940	0.878	0.705	0.235
	Llama-3.1-8B	0.898	0.798	0.575	0.951	0.850	0.575	0.376	0.850	0.772	0.660	0.889	0.809	0.660	0.229
	Mistral-7B-v0.3	0.898	0.828	0.629	0.949	0.852	0.629	0.320	0.853	0.798	0.740	0.879	0.832	0.740	0.139
	Qwen2.5-7B	0.913	0.847	0.644	0.963	0.876	0.644	0.319	0.887	0.832	0.721	0.926	0.855	0.721	0.205
	Average	0.905	0.829	0.630	0.956	0.867	0.630	0.326	0.872	0.806	0.707	0.908	0.844	0.707	0.202
NQ	Gemma-3-12B	0.862	0.852	0.692	0.940	0.832	0.692	0.248	0.845	0.838	0.763	0.892	0.825	0.763	0.129
	Llama-3.1-8B	0.828	0.786	0.538	0.919	0.773	0.538	0.381	0.788	0.786	0.708	0.814	0.741	0.708	0.106
	Mistral-7B-v0.3	0.815	0.778	0.564	0.904	0.768	0.564	0.340	0.766	0.766	0.735	0.778	0.747	0.735	0.043
	Qwen2.5-7B	0.851	0.812	0.575	0.934	0.789	0.575	0.359	0.807	0.798	0.727	0.838	0.772	0.727	0.111
	Average	0.839	0.807	0.592	0.924	0.791	0.592	0.332	0.801	0.797	0.733	0.830	0.771	0.733	0.097
Image QA															
AdvQA	Gemma-3-12B	0.777	0.702	0.502	0.874	0.766	0.502	0.372	0.740	0.670	0.558	0.805	0.733	0.558	0.247
	Pixtral-12B	0.761	0.646	0.446	0.878	0.745	0.446	0.432	0.677	0.552	0.541	0.719	0.706	0.541	0.179
	Qwen2.5-VL-7B	0.795	0.740	0.558	0.893	0.783	0.558	0.335	0.739	0.692	0.683	0.766	0.757	0.683	0.083
	Average	0.778	0.696	0.502	0.882	0.765	0.502	0.380	0.718	0.638	0.594	0.764	0.732	0.594	0.170
OK-VQA	Gemma-3-12B	0.849	0.821	0.675	0.929	0.841	0.675	0.254	0.839	0.816	0.746	0.884	0.833	0.746	0.138
	Pixtral-12B	0.823	0.765	0.560	0.918	0.796	0.560	0.358	0.784	0.751	0.650	0.841	0.760	0.650	0.191
	Qwen2.5-VL-7B	0.830	0.784	0.457	0.950	0.777	0.457	0.493	0.777	0.760	0.698	0.809	0.762	0.698	0.111
	Average	0.834	0.790	0.564	0.932	0.805	0.564	0.368	0.800	0.776	0.698	0.845	0.785	0.698	0.147
VQAv2	Gemma-3-12B	0.862	0.814	0.597	0.953	0.851	0.597	0.356	0.855	0.811	0.682	0.920	0.846	0.682	0.238
	Pixtral-12B	0.841	0.835	0.699	0.913	0.815	0.699	0.214	0.836	0.832	0.707	0.903	0.810	0.707	0.195
	Qwen2.5-VL-7B	0.833	0.755	0.469	0.940	0.806	0.469	0.471	0.766	0.748	0.738	0.778	0.772	0.738	0.040
	Average	0.845	0.801	0.588	0.935	0.824	0.588	0.347	0.819	0.797	0.709	0.867	0.809	0.709	0.158

Entropy matching improves ES-xAUC and reduces ES-δAUC in all 21 model–dataset combinations. The improvement is consistently driven by the entropy-misaligned comparison $\mathrm{AUC}(C_H, I_L)$, which is the minimum atomic AUROC in every reported setting.

Averaged over the text-QA datasets, $\mathrm{AUC}(C_H, I_L)$, and consequently ES-xAUC, increases from 62.2% to 71.7%, while ES-δAUC decreases from 31.2% to 14.9%. For image QA, ES-xAUC increases from 55.1% to 66.7%, and ES-δAUC decreases from 36.5% to 15.8%.

The improved balance across entropy subsets is accompanied by a moderate decrease in pooled correctness AUROC. CvI decreases from 86.5% to 83.1% for text QA and from 81.9% to 77.9% for image QA. This is expected: the natural probe benefits from the entropy–correctness association on the dominant entropy-aligned population, whereas matching prevents it from exploiting that signal. More importantly, a residual ES-δAUC of approximately 15% remains in both modalities. Removing entropy discrimination from the training distribution therefore mitigates, but does not completely eliminate, the shortcut.

A.7.4 Geometric interpretation

We next examine whether entropy matching changes the decision direction learned by the correctness probe. Let $\mathcal{D}(\alpha)$ denote a sequence of progressively matched training distributions, where $\mathcal{D}(0) =$ `Natural` and $\mathcal{D}(1) =$ `Matched`. We use $\alpha \in \{0, 0.25, 0.50, 0.75, 1\}$ and train an independent correctness probe with direction $\mathbf{w}_g(\alpha)$ at each matching level. The semantic entropy probe direction, $\mathbf{w}_{\mathrm{SEP}}$, is trained once on the natural split and held fixed throughout the comparison.

A smaller angle between these directions indicates that the correctness probe is more closely aligned with a probe explicitly trained to predict entropy. We report both a covariance-weighted functional

Table 13: Model-wise angle (degrees) between the fixed entropy direction w_{SEP}, learned on the natural training distribution, and the correctness direction $w_g(\alpha)$ learned at each matching level α, measured in the covariance-weighted (functional) metric θ_Σ and the Euclidean (parameter-space) metric θ_{E}. Here, $\alpha = 0$ denotes natural training and $\alpha = 1$ denotes fully entropy-matched training. $\Delta\theta = \theta(1) - \theta(0)$, computed from unrounded angles, so it may differ slightly from the difference of the displayed endpoints. Larger angles indicate greater separation between correctness and entropy. Light-red columns report the total angular change across the matching path.

Dataset	Model	Covariance-weighted angle θ_Σ						Euclidean angle θ_{E}					
		Matching level α					Change	Matching level α					Change
		0.00	**0.25**	**0.50**	**0.75**	**1.00**	$\Delta\theta_\Sigma$	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**	$\Delta\theta_{\mathrm{E}}$
LLM													
SciQ	Llama-3.1-8B	42.9	44.4	46.9	52.6	68.7	25.8	64.5	65.8	67.6	71.3	79.7	15.2
	Mistral-7B-v0.3	43.7	45.7	48.8	53.9	67.5	23.8	66.1	67.7	70.0	73.8	80.7	14.5
	Qwen2.5-7B	40.8	42.5	44.4	48.8	57.8	17.0	66.3	67.7	69.7	72.7	77.9	11.6
	Gemma-3-12B	40.8	41.3	42.5	43.8	47.0	6.2	70.5	71.5	73.0	74.9	78.5	8.0
TriviaQA	Llama-3.1-8B	22.8	24.2	26.1	30.2	48.4	25.6	50.9	52.8	56.1	61.8	76.9	26.0
	Mistral-7B-v0.3	26.0	27.5	30.2	35.4	52.9	26.9	51.7	54.0	57.4	63.2	76.6	24.9
	Qwen2.5-7B	22.7	24.5	27.0	30.8	41.2	18.5	49.3	52.4	56.5	61.1	70.5	21.2
	Gemma-3-12B	22.7	23.5	24.8	27.1	32.2	9.5	56.7	58.7	61.5	64.8	72.0	15.3
NQ	Llama-3.1-8B	38.9	42.5	47.2	54.5	66.9	28.1	60.0	63.3	67.3	72.3	79.7	19.8
	Mistral-7B-v0.3	43.6	47.6	52.7	60.5	74.1	30.4	60.9	63.6	67.4	72.3	79.5	18.5
	Qwen2.5-7B	35.6	38.3	42.3	49.0	61.0	25.3	58.5	61.3	64.7	69.7	77.2	18.7
	Gemma-3-12B	42.4	44.3	46.8	50.6	55.5	13.1	63.5	65.6	68.3	71.6	75.8	12.2
VLM													
AdvQA	Gemma-3-12B	39.4	41.7	44.8	49.1	54.3	14.9	79.5	81.5	84.2	87.7	90.8	11.4
	Pixtral-12B	40.5	42.9	48.2	56.0	68.1	27.6	59.1	62.4	67.0	73.5	81.9	22.7
	Qwen2.5-VL-7B	49.5	53.5	59.1	66.6	77.6	28.1	58.1	61.0	65.4	70.8	78.3	20.2
OKVQA	Gemma-3-12B	48.3	50.3	52.7	56.1	61.0	12.7	72.8	74.3	76.0	78.5	81.5	8.6
	Pixtral-12B	40.9	42.9	46.2	51.1	61.4	20.5	65.4	67.9	71.0	75.7	83.4	18.0
	Qwen2.5-VL-7B	44.5	47.0	50.7	57.6	73.2	28.7	67.9	69.2	71.2	74.7	81.9	14.0
VQAv2	Gemma-3-12B	34.2	35.7	37.7	40.7	45.8	11.5	69.1	71.3	73.9	77.1	82.0	12.9
	Pixtral-12B	44.1	44.9	45.7	46.8	48.0	3.8	74.6	76.1	77.9	79.9	82.6	8.0
	Qwen2.5-VL-7B	39.5	42.5	46.2	54.1	73.6	34.1	59.9	61.7	64.1	68.3	77.1	17.1

angle and the ordinary Euclidean angle using the common definition

$$\theta_{\mathbf{A}}(\alpha) = \cos^{-1}\left(\frac{\mathbf{w}_{\mathrm{SEP}}^\top \mathbf{A}\mathbf{w}_g(\alpha)}{\sqrt{\mathbf{w}_{\mathrm{SEP}}^\top \mathbf{A}\mathbf{w}_{\mathrm{SEP}}}\sqrt{\mathbf{w}_g(\alpha)^\top \mathbf{A}\mathbf{w}_g(\alpha)}}\right), \quad \mathbf{A} = \begin{cases} \widehat{\mathbf{\Sigma}}, & \theta_\Sigma, \\ \mathbf{I}, & \theta_{\mathrm{E}}, \end{cases} \tag{50}$$

where $\widehat{\mathbf{\Sigma}}$ is the covariance of the centered hidden representations from the fixed natural test set. The functional angle θ_Σ is equivalent to measuring the angle between the centered score vectors produced by the two probes on these reference examples. It therefore discounts parameter directions that exhibit little variation in the observed representations. The intercepts are omitted because centering removes constant score offsets.

Table 13 shows that both angles increase monotonically with the matching level for every model–dataset pair. From $\alpha = 0$ to $\alpha = 1$, the covariance-weighted angle increases by 20.9° on average for text QA and 20.2° for image QA. The corresponding average Euclidean increases are 17.2° and 14.8°, respectively. Because $\mathbf{w}_{\mathrm{SEP}}$, the covariance matrix, and the reference examples remain fixed across all matching levels, this trend can only arise from changes in the learned correctness direction.

The geometric and predictive results therefore provide complementary evidence. As the entropy–correctness association is progressively removed, the correctness probe rotates away from the semantic-entropy direction, its performance on C_H versus I_L improves, and the disparity among the four atomic comparisons decreases. Nevertheless, the remaining performance gap and the sample loss introduced by matching show that dataset manipulation alone is insufficient. The following sections therefore consider Joint Learning and Routed Learning, which mitigate the shortcut while training on the original `Natural` split.

Table 14: **Correctness prediction performance of the LLM probe trained using the proposed joint learning and routed learning schemes.** They are compared against several prior works on LLM probes. All evaluated prior LLM probes exhibit evidence of the entropy shortcut. The results are averaged over LLMs / Multimodal LLMs and model-wise results are in Appendix Table 15. CvI is the pooled correct-versus-incorrect AUROC.

Text QA													
Method	**Band AUC**	**AUC(C_H, I_L) ↑**			**ES-xAUC ↑**			**ES-δAUC ↓**			**CvI ↑**		
		NQ	**SciQ**	**TriviaQA**	**NQ**	**SciQ**	**TriviaQA**	**NQ**	**SciQ**	**TriviaQA**	**NQ**	**SciQ**	**TriviaQA**
Semantic Entropy		0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.750	0.728	0.830
SEP		0.280	0.394	0.484	0.280	0.394	0.484	0.648	0.514	0.466	0.754	0.777	0.875
Correctness Probe		0.592	0.643	0.630	0.592	0.643	0.630	0.332	0.279	0.326	**0.839**	**0.851**	**0.905**
SINK		0.556	0.597	0.567	0.556	0.597	0.567	0.348	0.294	0.379	0.811	0.816	0.885
LapEigVals		0.534	0.516	0.526	0.534	0.516	0.526	0.351	0.325	0.386	0.789	0.754	0.842
AttnLogDet		0.552	0.569	0.562	0.552	0.569	0.562	0.341	0.215	0.342	0.800	0.723	0.842
AttnEigVals		0.541	0.543	0.544	0.541	0.543	0.544	0.327	0.237	0.322	0.775	0.715	0.804
Joint Learning (DRO)		0.731	0.720	0.709	0.731	0.720	0.709	0.110	0.140	0.202	0.808	0.823	0.874
Routed Learning		**0.804**	**0.800**	**0.846**	**0.737**	**0.749**	**0.723**	**0.068**	**0.052**	**0.125**	0.761	0.774	0.764
Image QA													
Method	**Band AUC**	**AUC(C_H, I_L) ↑**			**ES-xAUC ↑**			**ES-δAUC ↓**			**CvI ↑**		
		VQAv2	**OK-VQA**	**AdvQA**	**VQAv2**	**OK-VQA**	**AdvQA**	**VQAv2**	**OK-VQA**	**AdvQA**	**VQAv2**	**OK-VQA**	**AdvQA**
Semantic Entropy		0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.701	0.707	0.707
SEP		0.339	0.258	0.265	0.339	0.258	0.265	0.589	0.682	0.645	0.769	0.750	0.723
Correctness Probe		0.588	0.564	0.502	0.588	0.564	0.502	0.347	0.368	0.380	**0.845**	**0.834**	**0.778**
SINK		0.542	0.487	0.497	0.542	0.487	0.497	0.366	0.414	0.362	0.809	0.786	0.755
LapEigVals		0.526	0.473	0.490	0.526	0.473	0.490	0.365	0.404	0.348	0.789	0.765	0.735
AttnLogDet		0.463	0.446	0.502	0.463	0.446	0.484	0.244	0.257	0.134	0.634	0.626	0.579
AttnEigVals		0.513	0.467	0.524	0.513	0.467	0.524	0.313	0.319	0.166	0.735	0.692	0.642
Joint Learning (DRO)		0.745	0.697	0.568	0.745	0.697	0.568	0.082	0.144	0.236	0.802	0.800	0.740
Routed Learning		**0.788**	**0.760**	**0.677**	**0.760**	**0.742**	**0.643**	**0.044**	**0.037**	**0.083**	0.784	0.766	0.698

Band-AUC order: Each cell is first averaged across models within each dataset and then across the three datasets in its panel. Colors indicate AUROC: 0.0 1.0

A.8 Semantic entropy implementation and binarization

For every input, we generate one greedy response at temperature $T = 0$ and $N = 10$ stochastic responses at $T = 1.0$. The greedy response is used to determine correctness, whereas only the stochastic responses are used to compute discrete semantic entropy (DSE) Kuhn et al. (2023). For open-ended QA, semantically equivalent sampled responses are grouped using natural-language inference (NLI). If the resulting semantic clusters have sizes $n_1, \ldots, n_K$, DSE is

$$\mathrm{DSE}(X) = -\sum_{c=1}^{K} p_c \ln p_c, \qquad p_c = \frac{n_c}{N}. \tag{51}$$

We use the natural logarithm and do not normalize DSE for the text- and image-QA benchmarks. Table 16 summarizes the complete generation and clustering configuration.

Medical classification tasks. The fine-tuned generative VLMs retain their language-model heads and emit class labels as text. Because the output space is closed, sampled labels can be grouped directly and no NLI model is required. Each input produces one greedy label and ten stochastic labels using constrained decoding, $T = 1.0$, and top-$p = 1.0$. For a task with K classes, we normalize the resulting entropy as

$$\mathrm{DSE}_{\mathrm{norm}}(X) = \frac{\mathrm{DSE}(X)}{\ln K}. \tag{52}$$

A.8.1 Semantic entropy binarization

The atomic subsets used throughout the paper require a binary low- or high-entropy label. Following the procedure used for Semantic Entropy Probes Han et al. (2024), we independently determine an entropy threshold for each model–dataset pair. The threshold is estimated from the `Natural` training split and then held fixed when assigning bands to the corresponding validation, test, and modified training splits.

Table 15: Model–dataset-level atomic AUROC results for the proposed Joint Learning and Routed Learning approaches across text and image QA. Band AUC reports the four atomic comparisons. ES-xAUC is their minimum, while ES-δAUC is their maximum-to-minimum gap. CvI is the pooled correct-versus-incorrect AUROC.

Dataset	Model	Joint Learning							Routed Learning						
		Band AUC				ES-xAUC	ES-δAUC	CvI	Band AUC				ES-xAUC	ES-δAUC	CvI
		$\frac{C_H}{I_H}$	$\frac{C_H}{I_L}$	$\frac{C_L}{I_H}$	$\frac{C_L}{I_L}$	↑	↓	↑	$\frac{C_H}{I_H}$	$\frac{C_H}{I_L}$	$\frac{C_L}{I_H}$	$\frac{C_L}{I_L}$	↑	↓	↑
Text QA															
SciQ	Gemma	0.792	0.723	0.901	0.851	0.723	0.178	0.857	0.768	0.850	0.746	0.832	0.746	0.104	0.806
	Llama	0.740	0.703	0.834	0.806	0.703	0.131	0.802	0.748	0.791	0.745	0.786	0.745	0.046	0.759
	Mistral	0.750	0.720	0.822	0.804	0.720	0.102	0.794	0.740	0.758	0.749	0.766	0.740	0.025	0.752
	Qwen	0.827	0.734	0.884	0.808	0.734	0.150	0.838	0.770	0.801	0.766	0.797	0.766	0.035	0.781
TriviaQA	Gemma	0.816	0.717	0.936	0.880	0.717	0.219	0.897	0.793	0.904	0.712	0.850	0.712	0.192	0.792
	Llama	0.777	0.654	0.895	0.812	0.654	0.241	0.855	0.784	0.775	0.740	0.740	0.740	0.044	0.748
	Mistral	0.813	0.727	0.895	0.831	0.727	0.168	0.863	0.798	0.813	0.735	0.763	0.735	0.077	0.752
	Qwen	0.828	0.736	0.915	0.854	0.736	0.179	0.881	0.819	0.892	0.707	0.822	0.707	0.185	0.764
NQ	Gemma	0.843	0.776	0.888	0.828	0.776	0.111	0.846	0.825	0.855	0.763	0.805	0.763	0.092	0.794
	Llama	0.781	0.709	0.820	0.756	0.709	0.111	0.791	0.778	0.771	0.722	0.716	0.716	0.062	0.744
	Mistral	0.771	0.707	0.819	0.762	0.707	0.112	0.786	0.763	0.770	0.737	0.743	0.737	0.032	0.749
	Qwen	0.799	0.733	0.838	0.777	0.733	0.104	0.808	0.800	0.819	0.732	0.757	0.732	0.087	0.756
Image QA															
AdvQA	Gemma	0.679	0.560	0.819	0.746	0.560	0.258	0.752	0.637	0.644	0.712	0.719	0.637	0.082	0.707
	Pixtral	0.570	0.526	0.747	0.707	0.526	0.222	0.690	0.584	0.656	0.635	0.701	0.584	0.117	0.654
	QwenV	0.739	0.619	0.848	0.763	0.619	0.228	0.778	0.708	0.731	0.732	0.757	0.708	0.049	0.733
OK-VQA	Gemma	0.808	0.782	0.844	0.824	0.782	0.062	0.825	0.781	0.804	0.798	0.820	0.781	0.040	0.809
	Pixtral	0.738	0.671	0.816	0.757	0.671	0.146	0.773	0.731	0.758	0.726	0.753	0.726	0.032	0.739
	QwenV	0.776	0.639	0.865	0.762	0.639	0.226	0.803	0.758	0.719	0.757	0.719	0.719	0.040	0.750
VQAv2	Gemma	0.780	0.758	0.838	0.828	0.758	0.080	0.822	0.753	0.779	0.801	0.826	0.753	0.073	0.808
	Pixtral	0.803	0.731	0.856	0.793	0.731	0.125	0.811	0.827	0.827	0.795	0.794	0.794	0.033	0.805
	QwenV	0.752	0.745	0.785	0.781	0.745	0.040	0.773	0.748	0.759	0.732	0.743	0.732	0.028	0.740

Within each Band-AUC block, the columns report $\mathrm{AUC}(C_H, I_H)$, $\mathrm{AUC}(C_H, I_L)$, $\mathrm{AUC}(C_L, I_H)$, and $\mathrm{AUC}(C_L, I_L)$, respectively. Light-red shading highlights the entropy-misaligned $\mathrm{AUC}(C_H, I_L)$ comparison and the two summary metrics. **Gemma**, **Llama**, **Mistral**, **Qwen**, **Pixtral**, and **QwenV** denote `Gemma-3-12B`, `Llama-3.1-8B`, `Mistral-7B-v0.3`, `Qwen2.5-7B`, `Pixtral-12B`, and `Qwen2.5-VL-7B`, respectively.

Table 16: Implementation details for discrete semantic entropy. The greedy generation is used to determine correctness, while the ten stochastic generations are used to compute DSE.

Component	Implementation
Generations	One greedy generation at $T = 0$ and ten stochastic generations at $T = 1.0$ per input.
Maximum new tokens	15 for text QA and 50 for image QA.
Few-shot examples	Five for text QA, sampled once per run using a fixed seed; none for image QA.
Top-p/top-k	Not overridden for the general QA experiments. We use the generation configuration packaged with each model checkpoint, e.g., top-$p = 0.9$ for Llama, top-$p = 0.8$ and top-$k = 20$ for Qwen, and top-$p = 0.95$ and top-$k = 64$ for Gemma-3.
NLI model	`microsoft/deberta-v2-xlarge-mnli` He et al. (2021). Its labels 0, 1, and 2 correspond to contradiction, neutral, and entailment, respectively.
Equivalence rule	Two answers are considered semantically equivalent only when the NLI model predicts entailment in both directions.
Clustering	Greedy semantic clustering Kuhn et al. (2023). The first unassigned answer initializes a cluster, and each subsequent answer joins the first cluster with which it is equivalent. Pairwise transitivity is not enforced.
Entropy	$-\sum_c p_c \ln p_c$, where $p_c = n_c/N$. We use the natural logarithm without normalization for the general QA tasks.
VLM inputs	The image and question are passed together through the processor's chat template for every generation. The generated answer texts are subsequently clustered using the same procedure as for text QA.

Table 17: Composition of the correctness–entropy subsets for the medical image-classification tasks. Here, H and L denote high and low sampling entropy, respectively. Percentages report each subset as a fraction of the complete evaluation set.

Dataset	Model	N	Count				% of N			
			$\lvert C_H \rvert$	$\lvert C_L \rvert$	$\lvert I_H \rvert$	$\lvert I_L \rvert$	C_H	C_L	I_H	I_L
Camelyon17	Qwen2-VL-2B	20,000	1,277	15,416	1,023	2,284	6.4	77.1	5.1	11.4
	SmolVLM2-2.2B	20,000	877	17,735	428	960	4.4	88.7	2.1	4.8
RSNA	Qwen2-VL-2B	2,668	1,138	315	1,017	198	42.7	11.8	38.1	7.4
	SmolVLM2-2.2B	2,668	978	803	650	237	36.7	30.1	24.4	8.9

C_H and I_L are entropy-misaligned subsets and are highlighted in light red; C_L and I_H are entropy-aligned. Percentages may not sum to exactly 100 due to rounding.

Let $s_j = \mathrm{DSE}(X_j)$. For a candidate threshold γ, define

$$\mathcal{S}_L(\gamma) = \{j : s_j \le \gamma\}, \qquad \mathcal{S}_H(\gamma) = \{j : s_j > \gamma\},$$
$$\mu_b(\gamma) = \frac{1}{|\mathcal{S}_b(\gamma)|} \sum_{j \in \mathcal{S}_b(\gamma)} s_j, \qquad b \in \{L, H\}. \tag{53}$$

We evaluate 100 evenly spaced candidate thresholds between zero and the maximum training-set DSE for that model–dataset pair. The selected threshold minimizes the total within-band squared deviation:

$$\gamma^\star = \arg\min_{\gamma \in \Gamma} \left[\sum_{j \in \mathcal{S}_L(\gamma)} \big(s_j - \mu_L(\gamma)\big)^2 + \sum_{j \in \mathcal{S}_H(\gamma)} \big(s_j - \mu_H(\gamma)\big)^2 \right]. \tag{54}$$

Candidates that leave either group empty are assigned infinite cost, and ties are resolved by selecting the first minimum. This procedure is a two-class Fisher–Jenks partition Fisher (1958), equivalently a one-dimensional two-means split evaluated over the candidate grid.

The entropy band of example j is then

$$B_j = \begin{cases} L, & s_j \le \gamma^\star, \\ H, & s_j > \gamma^\star. \end{cases} \tag{55}$$

Thus, an example lying exactly on the threshold belongs to the low-entropy band. Combining B_j with greedy-response correctness gives the four subsets C_L, C_H, I_L, and I_H. The binarization uses only semantic-entropy values and is not optimized using correctness labels. For the medical tasks, the same procedure is applied to $\mathrm{DSE}_{\mathrm{norm}}$.

A.9 Safety-critical applications

We additionally evaluate the entropy shortcut in two medical image-classification tasks: Camelyon17, which detects tumor tissue in lymph-node histopathology patches, and the RSNA Pneumonia Detection task, which classifies chest radiographs as normal, abnormal without lung opacity, or containing lung opacity. These tasks provide a safety-relevant test of correctness detection: a low-entropy incorrect prediction may evade uncertainty-based intervention, whereas a high-entropy correct prediction may trigger an unnecessary deferral.

Experimental setting. We fine-tune `Qwen2-VL-2B` and `SmolVLM2-2.2B` separately on each medical classification task. After fine-tuning, the VLM parameters are frozen and correctness probes are trained using their hidden representations. Evaluation uses 20,000 Camelyon17 images from the out-of-distribution validation split and 2,668 RSNA images from the held-out test split. The medical DSE computation is described in Appendix A.8, and the correctness-probe splits are reported in Appendix A.1.1.

Correctness–entropy composition. Table 17 shows that entropy-misaligned predictions remain substantial after task-specific fine-tuning, although their composition differs markedly between the two applications. On Camelyon17, C_L is the dominant subset, accounting for 77.1% and 88.7% of all

Table 18: Model-wise correctness-detection performance on the Camelyon17 and RSNA medical image-classification tasks. ES-xAUC is the minimum AUROC across the four atomic comparisons, while ES-δAUC is their maximum-to-minimum gap. CvI is the pooled correct-versus-incorrect AUROC.

Model	Method	CHvIH	CHvIL	CLvIH	CLvIL	ES-xAUC	ES-δAUC	CvI
Camelyon17								
Qwen2-VL-2B	Semantic Entropy	0.521	0.000	1.000	0.669	0.000	1.000	0.725
	SEP	0.565	0.173	0.971	0.812	0.173	0.798	0.818
	Correctness Probe	0.731	0.677	0.953	0.930	0.677	0.276	**0.919**
	Joint Learning	0.714	0.794	0.865	0.902	0.714	0.188	0.881
	Routed Learning	0.717	**0.833**	0.757	0.861	**0.717**	**0.144**	0.826
SmolVLM2-2.2B	Semantic Entropy	0.504	0.000	1.000	0.706	0.000	1.000	0.767
	SEP	0.519	0.133	0.985	0.875	0.133	0.853	0.878
	Correctness Probe	0.727	0.631	0.974	0.949	0.631	0.343	**0.943**
	Joint Learning	0.705	0.773	0.891	0.919	**0.705**	**0.214**	0.903
	Routed Learning	0.677	**0.891**	0.687	0.881	0.677	**0.214**	0.821
RSNA								
Qwen2-VL-2B	Semantic Entropy	0.508	0.027	0.986	0.595	0.027	0.958	0.537
	SEP	0.532	0.278	0.838	0.623	0.278	0.559	0.559
	Correctness Probe	0.548	0.403	0.752	0.622	0.403	0.349	**0.570**
	Joint Learning	0.539	0.479	0.650	0.593	0.479	0.172	0.553
	Routed Learning	0.549	**0.502**	0.622	0.570	**0.502**	**0.119**	0.557
SmolVLM2-2.2B	Semantic Entropy	0.473	0.022	0.984	0.631	0.022	0.962	0.595
	SEP	0.523	0.259	0.865	0.681	0.259	0.606	0.617
	Correctness Probe	0.559	0.391	0.809	0.669	0.391	0.418	**0.630**
	Joint Learning	0.548	0.496	0.664	0.612	0.496	0.168	0.587
	Routed Learning	0.539	**0.558**	0.601	0.621	**0.539**	**0.082**	0.572

predictions from Qwen2-VL-2B and SmolVLM2-2.2B, respectively. Nevertheless, most remaining errors have low entropy: I_L constitutes 69.1% of Qwen2-VL-2B errors and 69.2% of SmolVLM2-2.2B errors. These confidently repeated errors are precisely the cases that an entropy-based detector is least likely to flag.

RSNA exhibits a different form of misalignment. Here, C_H alone accounts for 42.7% and 36.7% of the complete evaluation set for the two models. Equivalently, 78.3% of Qwen2-VL-2B's correct predictions and 54.9% of SmolVLM2-2.2B's correct predictions lie in the high-entropy band. Overall, the two entropy-misaligned subsets, $C_H \cup I_L$, comprise 50.1% and 45.6% of the RSNA predictions. Thus, the failure of the usual inverse entropy–correctness association is not restricted to open-ended QA and can manifest differently across medical tasks.

Correctness-detection performance. Table 18 reports the complete model-wise atomic AUROCs. Semantic entropy exhibits the expected extreme shortcut: $\mathrm{AUC}(C_H, I_L)$ is between 0.000 and 0.027, with ES-δAUC between 0.958 and 1.000. SEP partially improves this comparison but remains strongly entropy-aligned. The standard Correctness Probe performs considerably better, yet $\mathrm{AUC}(C_H, I_L)$ remains its weakest atomic comparison for all four model–dataset pairs.

On Camelyon17, the model-averaged Correctness Probe obtains an ES-xAUC of 65.4% and an ES-δAUC of 31.0%. Joint Learning increases ES-xAUC to 71.0% and reduces the gap to 20.1%. Routed Learning produces the strongest entropy-misaligned comparison, $\mathrm{AUC}(C_H, I_L) = 86.2\%$, and the smallest disparity, 17.9%, while attaining an ES-xAUC of 69.7%.

The gains are larger on RSNA. Relative to the Correctness Probe, Joint Learning increases the model-averaged ES-xAUC from 39.7% to 48.7% and reduces ES-δAUC from 38.3% to 17.0%. Routed Learning further improves these metrics to 52.1% and 10.1%, respectively. Its advantage also holds for both individual VLMs, demonstrating that the result is not driven by averaging.

These findings extend the entropy shortcut beyond general QA: task-specific fine-tuning does not prevent medical VLM correctness probes from exploiting the entropy–correctness association. Joint and Routed Learning substantially improve stability across entropy regimes, particularly for low-entropy errors. These benchmark results demonstrate robustness in safety-relevant settings, but should not be interpreted as establishing clinical readiness.



Figure 5: **Examples of four entropy-correctness associations from Text QA task**: The figure shows examples (from left to right columns) of low-entropy correct, high-entropy correct, low-entropy incorrect, and high-entropy incorrect samples from different datasets.

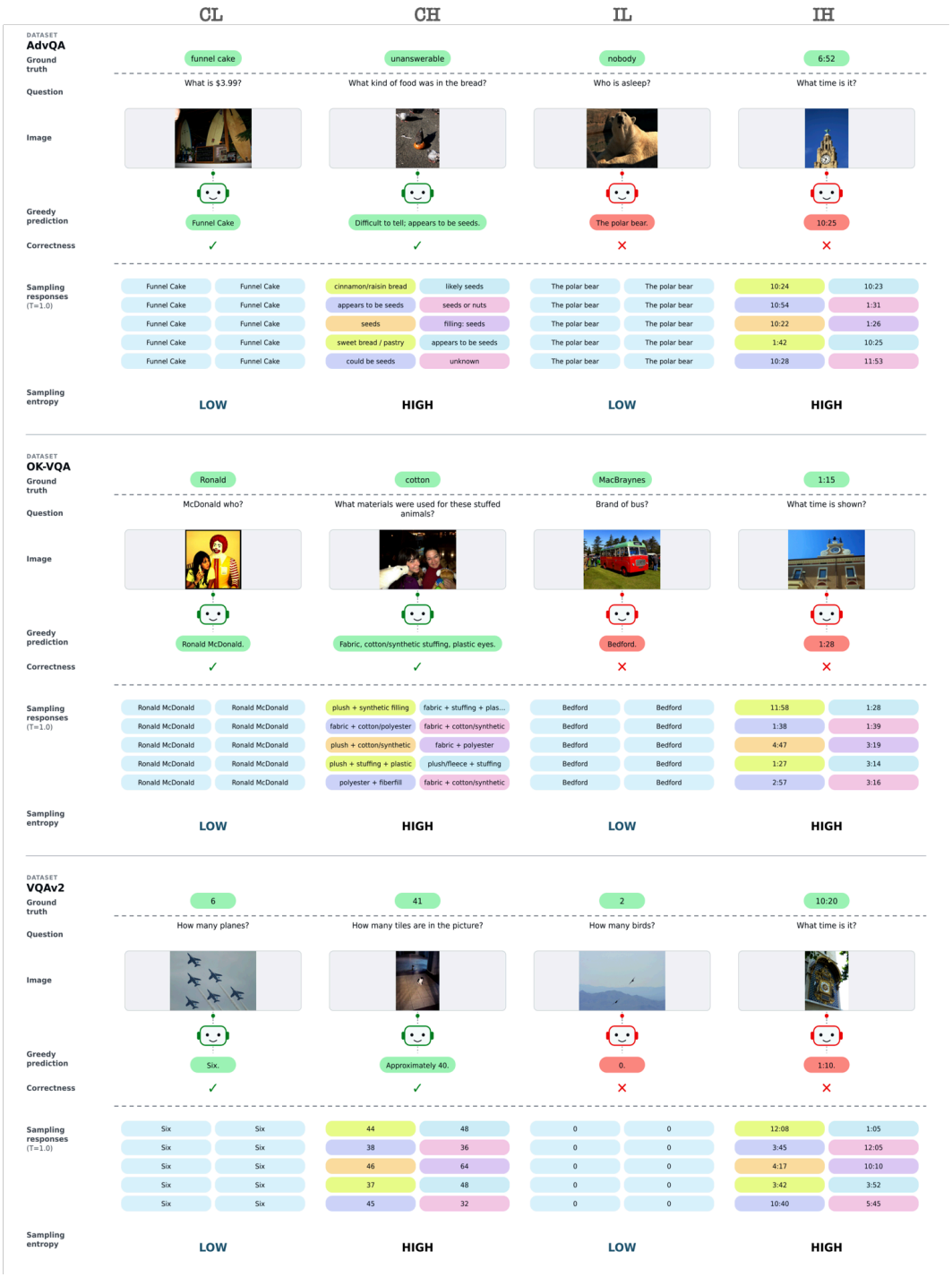


Figure 6: **Examples of four entropy-correctness associations from Image QA task**: The figure shows examples (from left to right columns) of low-entropy correct, high-entropy correct, low-entropy incorrect, and high-entropy incorrect samples from different datasets.