ESTIMATING UNCERTAINTY OF OMNIMODAL LARGE LANGUAGE MODELS VIA EXTERNAL OMNIMODAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has investigated Omni-modal Large Language Models (OLLMs), models which can take multiple modalities and their combinations as input, as a way forward to have systems really capable of interacting with their environment. Uncertainty quantification (UQ) for OLLMs, however, remains an open problem: existing methods largely require access to model internals, or expensive self-verbalisation approaches. We propose a model-agnostic UQ framework that scores a Chain-of-Thought response entirely through external omni-modal embeddings, measuring the *grounding* of the reasoning in the multimodal input and its internal *coherence*. The resulting confidence estimates require no access to logits, hidden states or gradients of the evaluated model, rather relying on the segmented reasoning trace and, for the aggregated score, a labelled calibration set that we show can be as small as 25 examples. On UNO-Bench, with both a closed and an open model, our scores yield small gains over self-consistency (systematic with the closed generator, mixed with the open one) and are competitive with self-verbalisation and internal-state baselines, especially in open-ended, omni-modal questions where previous state-of-the-art baselines struggle.[1]

1 INTRODUCTION

Omni-modal large language models (OLLMs) process and reason over diverse modalities including text, images, audio and video. Assessing the reliability of their outputs through Uncertainty Quantification (UQ) is an open problem.

Recent attempts at UQ for multimodal models (Lau et al., 2026) exhibit major limitations: (1) they require access to model internal states or final logits, limiting their scope to open-source models; (2) they are tested primarily on bimodal interactions (e.g. audio-text) instead of omni-modal (e.g. audio, image and text at the same time); and (3) they typically rely solely on the final answer, ignoring the information contained in the model's reasoning trace or Chain-of-Thought (CoT).

To address these issues, we introduce a novel model-agnostic UQ method for omni-modal contexts, which derives its confidence signal from the geometry of the model's reasoning trace in the shared space of omni-modal embedding models external to the analysed model. Our approach maps to a shared embedding space both the multi-modal inputs, the reasoning steps and the final answer, deriving *grounding* and *coherence* scores based on the distance between such embeddings. Such scores are then used to grade the majority-selected answer so that ungrounded and low-coherence answers get less *voting power*. We find small but statistically systematic gains over standard self-consistency across two answer-generating models (a closed and an open one), which are largest exactly in the two regimes that matter most for costly omni-modal APIs: low sampling budgets and open-ended questions spanning more than two input modalities at once, as answered by a close-source OLLM.

Our family of models reach the best average discrimination performance among all other baselines under both the close and open-source generators, while answer dispersion approaches retain an advantage in error calibration and in the Audio modality.

[1] Code: https://anonymous.4open.science/r/Multimodal-CoT-Confidence-Scoring-C12E

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.

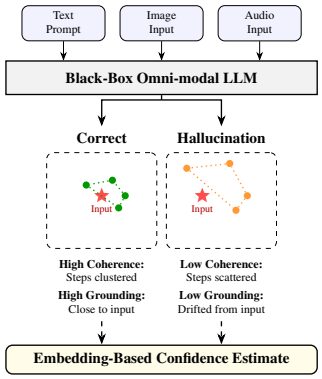


Figure 1: We map the omni-modal inputs and generated Chain-of-Thought reasoning steps into a shared semantic space to compute **Coherence** and **Grounding**, which grade the majority-selected answer without internal model logits and without any additional call to the evaluated model.

Our scores are nonetheless markedly more interpretable than self-consistency and, especially, previous embedding-based methods relying on internal states and answer-only embeddings, since they have a straightforward geometric interpretation rooted in well-defined individual scores. Finally, we release both CoT datasets used in our experiments, derived from answering the questions of an indicative benchmark with a closed and an open-source model respectively. We hope that by doing so we will be able to foster research in this interesting and still relatively under-developed area of research.

2 Related Work

2.1 LLM Uncertainty Quantification

Literature on uncertainty quantification for LLMs is abundant, starting from the seminal work by Wang et al. (2022) showing that self-consistency is a good signal of LLMs certainty over their output. Most works in this context, then, have developed further the self-consistency direction with better answer matching (Farquhar et al., 2024). At the same time, other work has shown that the log-probabilities of the models could also serve as a strong signal of model (un)certainty (Fadeeva et al., 2024), even though its alignment to true probabilities has been questioned (Toney & Wails, 2025). Similarly, another line of work has explored the internal states of the LLMs as a strong signal for uncertainty measure (Bu et al., 2026). The preponderance of private models, for which neither log probabilities nor hidden states are available, has however led much of the literature to focus on black-box, external methods such as self-consistency. One such method consists in asking the LLMs themselves to *verbalize* their confidence: such methods tend to work well for powerful models, but they fail when the LLM is not powerful enough and they tend to be overconfident (Mei et al., 2026).

2.2 Uncertainty Quantification for Multimodal LLMs

Recent work has started exploring the space of multimodal LLMs, starting from Visual LLMs (Zhang et al., 2024). The biggest challenge in this context is potential conflict between modalities (Bezirganyan et al., 2025). However, current methods mostly follow approaches such as self-verbalisation, collapsing the role of different modalities into a single judgment (Chen et al., 2025b). A different approach, UMPIRE (Lau et al., 2026), was recently proposed which instead uses the internal states of the LLM to adjust Semantic Volume. Also this method, however, presents the problem of needing access to some model internal states. In case of private models, the authors circumvented this problem by feeding the input and answer to an open source multimodal LLM, but also this solution can be seen as suboptimal since it involves recomputing the entire states for a second model and it is contingent on how *similar* to the private model the open source model is.

3 METHODOLOGY

To address the limitations of prior work, we introduce a fully external framework that assesses the confidence of a generated answer using self-consistenty and the *coherence* and *grounding* of its reasoning chain, measured in the embedding space of an external encoder. The resulting scores need no access to the logits, hidden states or gradients the evaluated model and no calls to another (or the same) judging model; segmenting the trace into steps is done once by a separate auxiliary model, and the aggregated score assumes a small labelled calibration sample.

Specifically, we define *coherence* and *grounding* in such omni-modal embedding space as:

1. *grounding*: the semantic alignment, in the shared embedding space, between each intermediate reasoning step and the non-textual multi-modal inputs.
2. *coherence*: the dispersion of the embeddings of the steps of the generated reasoning chain (based on text only), which we further decompose into internal reasoning smoothness, density and convergence toward the predicted answer.

3.1 PROBLEM FORMULATION

Let $\mathcal{M}$ denote a blackbox omni-modal model, which receives a multimodal context

$$\mathcal{V} = \bigcup_{k=1}^{K} V^{(k)}, \quad V^{(k)} = \{v_1^{(k)}, \ldots, v_{n_k}^{(k)}\},$$

where each modality-specific subset $V^{(k)}$ collects the elements of one input modality. In our instantiation the pool contains only the non-textual inputs of the question (audio, image and video); the textual prompt is deliberately excluded from $\mathcal{V}$, so that the grounding scores of Section 3.2 cannot be satisfied by aligning a reasoning step with the question text. We assume access to the chain-of-thought process of the model, thus we define the model-generated reasoning chains

$$R = [r_1, r_2, \ldots, r_m]$$

followed by a final answer a.

Because internal representations of $\mathcal{M}$ might be unavailable, we instead rely on an external embedding model

$$\mathcal{E} : x \to \phi_x \in \mathbb{R}^d.$$

We embed all components into a shared space:

$$\Phi_{\mathcal{V}} = \{\phi_v : v \in \mathcal{V}\}, \quad \Phi_R = [\phi_{r_1}, \ldots, \phi_{r_m}], \quad \phi_a.$$

Let $\mathcal{S}(\cdot, \cdot)$ denote the normalized cosine similarity defined as

$$\mathcal{S}(x, y) = \frac{1 + \cos(\phi_x, \phi_y)}{2},$$

which maps similarity scores to the interval $[0, 1]$.

3.2 GROUNDING

We quantify grounding by measuring the alignment between reasoning steps and input modalities in a shared omni-modal embedding space.

Recall that the input context $\mathcal{V} = \bigcup_{k=1}^{K} V^{(k)}$ is partitioned into the K modality-specific subsets introduced above. The embedding model $\mathcal{E}$ maps each subset to its set of embeddings in the shared space:

$$\Phi^{(k)} = \{\phi_v : v \in V^{(k)}\}, \quad k = 1, \ldots, K. \tag{1}$$

Definition 1 (Omni-modal Grounding). *For a reasoning step r_i with embedding $\phi_{r_i} = \mathcal{E}(r_i)$, its alignment to modality k is defined as:*

$$A^{(k)}(r_i) = \max_{\phi_v \in \Phi^{(k)}} \mathcal{S}(\phi_{r_i}, \phi_v), \tag{2}$$

where $\mathcal{S}(\cdot, \cdot)$ is the normalized cosine similarity defined above.

Since we expect an individual reasoning step to often refer to one modality at a time, we then obtain the omni-modal alignment $A^{\text{omni}}(r_i)$ by applying the max operator again, this time over the marginal alignments $A^{(k)}(r_i)$:

$$A^{\text{omni}}(r_i) = \max_{k=1,\dots,K} A^{(k)}(r_i) \tag{3}$$

Then, having the alignments for each reasoning step i, we can aggregate with simple averaging:

$$G_{avg} = \frac{1}{m} \sum_{i=1}^{m} A^{\text{omni}}(r_i) \tag{4}$$

We further consider the max variant to factor in the most grounded step.

$$G_{\max} = \max_{i=1,\dots,m} A^{\text{omni}}(r_i) \tag{5}$$

Answer-level grounding. We additionally embed each candidate generation as a single unit and align it with the same input pool, giving three further scores $G^{\max}_{\text{ans}}$, $G^{\text{avg}}_{\text{ans}}$ and $G^{\min}_{\text{ans}}$ (Appendix G, Equation 16); the last requires support from *every* present modality. They cost one embedding comparison per candidate and no extra model call.

3.3 Answer Coherence

A *coherent* answer should exhibit gradual semantic transitions rather than abrupt jumps. We capture this property using three complementary geometric measures measured directly from the textual output of the OLLM (i.e. the reasoning chain).

Local Smoothness This quantity measures the local transition between one reasoning step to the next one by comparing pairwise, consecutive steps in the embedding space. Specifically, we compute:

$$S_{\text{smooth}} = \frac{1}{m-1} \sum_{i=1}^{m-1} \mathcal{S}(\phi_{r_i}, \phi_{r_{i+1}}) \tag{6}$$

Low smoothness indicates erratic reasoning transitions, highlighting low local coherence and possible hallucinations. Equations 6 and 7, and Δ_{bwd} below, are undefined for single-step chains; our implementation then assigns all three scores their maximum value 1. Such chains are rare with the closed generator (0.07% of candidates) but not with the open one (16.5%), whose coherence scores are inflated accordingly (Appendix G).

Semantic Density Semantic density measures the global coherence of the reasoning chain by comparing each step embeddings with each other and normalizing:

$$S_{\text{dens}} = \frac{1}{m(m-1)} \sum_{i \neq j} \mathcal{S}(\phi_{r_i}, \phi_{r_j}) \tag{7}$$

A low density in the embedding space indicates low global compactness within the reasoning chain, which has been shown to correlate with lower coherence (Ghinassi et al., 2024).

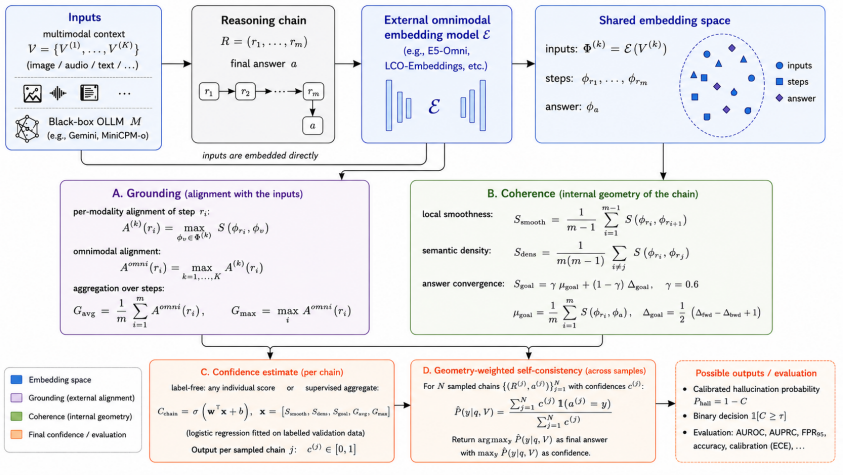


Figure 2: Visual overview of the proposed method, as explained in Methodology section.

Answer Convergence A reliable reasoning trajectory should also progressively move toward the final answer, a property also of a text coherently oriented towards a goal.

Let

$$\mu_{\text{goal}} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{S}(\phi_{r_i}, \phi_a)$$

denote the average proximity of reasoning steps to the answer embedding.

To capture directional progression, we define forward and backward alignments as

$$\Delta_{\text{fwd}} = \mathbb{E}_{i>m/2}[\mathcal{S}(\phi_{r_i}, \phi_a)] \tag{8}$$

$$\Delta_{\text{bwd}} = \mathbb{E}_{i\leq m/2}[\mathcal{S}(\phi_{r_i}, \phi_a)] \tag{9}$$

and compute the directional progression score as

$$\Delta_{\text{goal}} = \frac{1}{2}\left(\Delta_{\text{fwd}} - \Delta_{\text{bwd}} + 1\right). \tag{10}$$

which measures whether later reasoning steps move closer to the answer.

The final convergence score is

$$S_{\text{goal}} = \gamma\mu_{\text{goal}} + (1-\gamma)\Delta_{\text{goal}} \tag{11}$$

where γ is a weighting hyperparameter, fixed a priori to $\gamma = 0.6$ in all experiments.

3.4 Unified Chain-Level Score

The extracted chain-level signals

$$\mathbf{x} = [S_{\text{smooth}}, S_{\text{dens}}, S_{\text{goal}}, G_{\text{avg}}, G_{\text{max}}] \tag{12}$$

are used as feature inputs to an ℓ_1-regularised logistic regression classifier. The classifier is fitted on the validation set using answer correctness as the target label, learning the relationship between these

signals and the probability that a generated answer is correct. Specifically, the predicted correctness probability is computed as

$$\mathcal{C}_{\text{chain}} = \sigma\left(\mathbf{w}^\top \mathbf{x} + b\right), \tag{13}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\mathbf{w}$ and b are learned parameters. Fitting has two stages: C is chosen by grid search on validation AUROC, and the fitted model is then compared against each single-feature scorer on the same items, a single feature replacing it whenever strictly better. That explicit selection step, not the ℓ_1 penalty, is what yields the single-feature unit-weight solutions of Appendix F, so $\mathcal{C}_{\text{chain}}$ is best understood as validation-based selection of the strongest signal with a logistic-regression fallback rather than a genuinely multivariate combination (Appendix G).

We also report $\mathcal{C}^{+}_{\text{chain}}$, fitted with the identical protocol on Equation 12 extended by the three answer-level grounding scores of Equation 16 (Appendix G).

The final uncertainty estimate is then defined as the complement of the predicted correctness probability:

$$\mathcal{U} = 1 - \mathcal{C}_{\text{chain}}. \tag{14}$$

3.5 Geometry-Weighted Aggregation of Sampled Answers

Standard self-consistency treats all sampled reasoning chains equally: given N sampled generations $\{(R^{(j)}, a^{(j)})\}_{j=1}^{N}$, it estimates $\hat{P}(y|q,\mathcal{V}) = \frac{1}{N}\sum_j \mathbf{1}(a^{(j)} = y)$ which is treated as the model confidence. We instead compute for each sampled chain the feature vector $\mathbf{x}^{(j)}$ of Equation 12 and a chain-level confidence $c^{(j)}$ (either from the logistic-regression aggregator, $c^{(j)} = \mathcal{C}_{\text{chain}}(\mathbf{x}^{(j)})$, or directly from a single score in the individual-score variants) and replace the vote frequencies by their geometry-weighted counterpart:

$$\hat{P}(y|q,\mathcal{V}) = \frac{\sum_{j=1}^{N} c^{(j)} \mathbf{1}(a^{(j)} = y)}{\sum_{j=1}^{N} c^{(j)}}. \tag{15}$$

Equation 15 admits two modes, kept separate throughout. In the *primary* protocol of the main tables the returned answer is the unweighted count-majority cluster and the geometry only attaches a confidence to it, so every method grades the same answer. In the *active* mode (Appendix K) the answer is $\arg\max_y \hat{P}(y|q,\mathcal{V})$, so the geometry can change the prediction. The overall pipeline of the proposed framework is summarized in Figure 2. Unlike previous embedding-based methods, our procedure can be applied to any model, private or open, as it requires only externally observable artifacts: a generated reasoning chain and its final answer, embedded with an external encoder.

4 Experimental Setup

Dataset & Models: We use the uni- and omni-modal benchmark UNO-bench (Chen et al., 2025a), and report results on its four subsets corresponding to Audio only (Audio, 1,000 questions), Vision only (Visual, 1,480), Audio + Vision in a multiple choice answer setting (MC, 1,000) and Audio + Vision in an open-ended answer setting (MO, 250), together with their average (Avg).

We employ Gemini 3.1 Flash Lite and MiniCPM-o 4.5 (Cui et al., 2026) to generate 6 answers for each question in the UNO-bench dataset: our methods, Self-Consistency, and UMPIRE aggregate over candidates, whereas the self-verbalisation baselines score the selected answer through one additional call to the evaluated model. For the UMPIRE baseline with the closed generator, we extract internal states using MiniCPM-o 4.5 as a proxy model; in-depth experimental details are documented in Appendix A.

Embedding Model: We used E5-Omni (Chen et al., 2026) to generate embeddings. Appendix I repeats the evaluation with a different omnimodal encoder, LCO-Embedding-Omni-7B (Xiao et al., 2025), showing that our results are robust to this choice.

Answer Selection: Following Farquhar et al. (2024), sampled answers are grouped into semantic clusters by an LLM, and the candidate belonging to the most frequent cluster is selected with the methods scores breaking ties. The tie-break is therefore method-dependent, so on tied questions different methods can return different answers. Ties are not rare (28.9% of questions with Gemini, 14.2% with MiniCPM); per-split counts are in Appendix G. A variant in which the geometric evidence also *selects* the answer, via the weighted vote of Section 3.5, is evaluated in Appendix K.

Baselines: We compare against two families of methods. The first requires repeated sampling from the evaluated model: *self-consistency*, which uses the raw frequency of the selected answer's cluster among the $n = 6$ candidates as its confidence, and *UMPIRE*, which additionally requires access to internal states (extracted through a proxy model for closed-source generators). To the same sampling-based family belong two embedding-dispersion baselines computed from the candidate-*answer* embeddings of the same external encoder used by our method: *RDS* (radial dispersion score), the mean Euclidean distance of the N candidate-answer embeddings from their centroid, and *Semantic Volume*, the log-determinant of the Gram matrix of the mean-centred answer embeddings, $\log\det(XX^\top + \varepsilon I)$. The second family avoids repeated sampling but requires one additional call to the evaluated model: three SOTA self-verbalisation baselines. The first, which we call Level-Based, asks the model to assign a five-level confidence score to its answer, mapped to numerical values. The second and third follow Self-Probing (Liu et al., 2024) and Two-Step CoT (Tian et al., 2023), respectively, where the model estimates answer correctness from the question-answer pair and previously generated reasoning context. All baseline implementation details are in Appendix B; Appendix N analyses the dispersion baselines and further compares against two cluster-based baselines (discrete Semantic Entropy and the number of semantic sets) computed with the same clustering step.

Metrics: We evaluate Area Under the Receiver Operating Characteristic curve (AUROC) and Expected Calibration Error (ECE). Since the compared methods produce confidence scores on heterogeneous scales, before computing the metrics each method's scores are min–max rescaled to $[0, 1]$ over the evaluated split; this preprocessing leaves the rank-based AUROC unchanged and makes ECE comparable across methods. It is computed on the evaluated split itself, so it reflects range-normalised spread rather than deployment calibration; Appendix C reports Platt-calibrated ECE, where most of the large gaps disappear. Appendix E additionally reports AURAC metric. The main tables report mean values across 100 randomised runs. Statistical support is reported separately, in two complementary forms. First, the paired delta Table 3 tests the difference between $\mathcal{C}_{\text{chain}}$ and Self-Consistency on each split with a paired t-test. Second, Section 5.1 reports targeted paired-bootstrap comparisons over test questions on a canonical partition, which quantify generalisation to new questions; these are reported with their raw p-values.

5 RESULTS

We compare our individual scores and the aggregated $\mathcal{C}_{\text{chain}}$ against the sampling-based methods (Self-Consistency, UMPIRE, and the RDS and Semantic Volume answer-dispersion baselines) and the three self-verbalisation baselines; every method scores the same majority-selected answer (Section 4), so only the attached confidence differs, and our scores add one external embedding pass and no call to the evaluated model beyond the sampled candidates. MC and MO denote the OmniModal-Multiple Choice and OmniModal-Open Answer splits, respectively. Table 1 reports AUROC and ECE with Gemini-generated answers and Table 2 the same metrics with MiniCPM-generated answers; the Avg column is the unweighted mean of the four per-split values. Table 3 summarises the paired differences of our best methods against Self-Consistency across the 100 runs.

5.1 ITEM-LEVEL COMPARISONS

The run-level tests of Table 3 quantify robustness to evaluation randomness; we complement them with paired bootstrap deltas over test questions on a canonical partition ($B = 10^4$; Appendix C reports the underlying variances and confidence intervals). These comparisons are computed on the pooled union of the splits. With Gemini-generated answers, $\mathcal{C}_{\text{chain}}$ improves over Self-Consistency on pooled AUROC ($+0.016$, 95% CI $[+0.006, +0.025]$, $p = 0.0009$); with MiniCPM-generated answers it again improves over Self-Consistency (pooled AUROC $+0.014$, $p = 0.008$), whereas native UMPIRE holds a significant advantage over it (AUROC -0.035, $p = 0.002$). Against the strongest verbalised baseline, S_{dens} gains $+0.108$ AUROC over Self-Probing on Visual with Gemini

	Avg		Audio		MC		MO		Visual	
Method	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE
Self-Consistency	0.5927	0.3068	0.4512	0.6203	**0.5630**	0.2723	0.6334	0.2459	0.7234	**0.0888**
UMPIRE	0.5825	0.4507	0.6318	0.7075	0.5551	0.3999	0.6159	0.1307	0.5272	0.5649
RDS	0.6038	0.2405	0.6654	0.2760	0.5434	**0.1382**	0.5894	0.1951	0.6169	0.3526
Semantic Volume	0.6017	**0.2211**	**0.6671**	0.4689	0.5417	0.1802	0.5918	**0.1223**	0.6062	0.1130
Level-Based Conf.	0.5230	0.3192	0.5074	0.1661	0.5154	0.3466	0.5124	0.5777	0.5566	0.1866
Self-Probing Conf.	0.5803	0.3170	0.5705	**0.1594**	0.5495	0.3378	0.5871	0.5688	0.6143	0.2019
Two-Step CoT Conf.	0.5435	0.3147	0.5441	0.1611	0.5109	0.3473	0.5411	0.5522	0.5779	0.1981
S_{smooth}	0.6091	0.2970	0.4551	0.6124	0.5621	0.2758	**0.6880**	0.2071	0.7311	0.0928
S_{dens}	0.5923	0.3057	0.4355	0.6174	0.5405	0.2860	0.6563	0.2271	**0.7368**	0.0923
S_{goal}	0.5948	0.3077	0.4516	0.6141	0.5460	0.2895	0.6462	0.2302	0.7356	0.0969
G_{avg}	**0.6102**	0.2998	0.4817	0.6032	0.5576	0.2816	0.6705	0.2183	0.7311	0.0961
$G_{\max}$	0.5918	0.3043	0.4351	0.6148	0.5582	0.2718	0.6507	0.2355	0.7232	0.0952
$\mathcal{C}_{\text{chain}}$	0.6046	0.2999	0.4700	0.6068	0.5622	0.2752	0.6590	0.2236	0.7272	0.0939

Table 1: AUROC (↑) and ECE (↓) across the UNO-Bench splits with Gemini-generated answers (means over 100 randomised runs). Avg is the unweighted mean of the four per-split values; results on the pooled union of the splits are deferred to Appendix D, which explains why pooled AUROC is unreliable under split heterogeneity. Bold indicates the best result per column.

	Avg		Audio		MC		MO		Visual	
Method	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE
Self-Consistency	0.6559	0.2553	0.7236	0.1310	0.5461	0.3198	0.6782	0.4051	0.6755	0.1653
UMPIRE	0.6570	0.1577	**0.7507**	0.1550	0.5840	0.2898	0.5875	**0.1323**	**0.7058**	**0.0538**
RDS	0.5764	**0.1401**	0.7117	**0.1127**	0.4976	0.1664	0.5004	0.2133	0.5958	0.0680
Semantic Volume	0.5882	0.2027	0.7169	0.2858	0.5137	**0.1142**	0.5135	0.1770	0.6088	0.2338
Level-Based Conf.	0.6200	0.2699	0.6306	0.1796	0.6024	0.3905	0.5956	0.2745	0.6512	0.2350
Self-Probing Conf.	0.6563	0.2843	0.6941	0.1753	**0.6331**	0.4658	0.5933	0.2521	0.7047	0.2440
Two-Step CoT Conf.	0.6294	0.3133	0.6791	0.1830	0.6252	0.4804	0.5609	0.3028	0.6525	0.2871
S_{smooth}	0.6564	0.2628	0.7270	0.1406	0.5628	0.3116	0.6531	0.4357	0.6829	0.1632
S_{dens}	0.6538	0.2590	0.7344	0.1293	0.5572	0.3141	0.6347	0.4317	0.6887	0.1610
S_{goal}	0.6522	0.2627	0.7295	0.1354	0.5451	0.3155	0.6462	0.4408	0.6878	0.1593
G_{avg}	0.6588	0.2649	0.7419	0.1346	0.5507	0.3179	0.6675	0.4382	0.6750	0.1690
$G_{\max}$	0.6584	0.2564	0.7223	0.1332	0.5439	0.3225	**0.6978**	0.4013	0.6695	0.1686
$\mathcal{C}_{\text{chain}}$	**0.6603**	0.2573	0.7368	0.1330	0.5507	0.3178	0.6700	0.4161	0.6835	0.1623

Table 2: AUROC (↑) and ECE (↓) across the UNO-Bench splits with MiniCPM-generated answers. Avg and bold as in Table 1.

($p = 0.0001$) and S_{smooth} gains $+0.122$ AUROC on MO ($p = 0.009$); with MiniCPM, Self-Probing holds a significant pooled AUROC advantage (-0.039, $p = 0.003$) while our scores keep the lead on the open-answer omni-modal split (S_{goal}: $+0.134$ AUROC, $p = 0.023$).

5.2 Discussion

Several patterns emerge. First, the embedding scores are competitive with the agreement-based methods. With Gemini, G_{avg} attains the best average AUROC in Table 1 (0.6102), and the paired analyses show that $\mathcal{C}_{\text{chain}}$'s gains over Self-Consistency are systematic on three of the four splits with Gemini (average $+2.15\%$, Table 3) and hold at the item level under both generators (Section 5.1). With MiniCPM the gains are smaller but significant on three of the four splits (Audio $+1.82\%$, MC $+0.85\%$, Visual $+1.19\%$, average $+0.66\%$); the open-answer omni-modal split remains a loss (-1.20%), so we do not claim a uniform improvement across generators.

Second, the dispersion baselines are strong but their strength is concentrated where the agreement signal fails: both repair the Audio failure mode that inverts every agreement-based signal (AUROC 0.6654–0.6671, where Self-Consistency sits at 0.4512), and Semantic Volume attains the best average ECE in Table 1 (0.2211), yet on average AUROC they trail the best embedding scores (0.6017–0.6038 vs. 0.6102 for G_{avg}). The two signal families are thus complementary rather than redundant, consistently across generators: the dispersion baselines weaken exactly where the chain scores are strongest, which retain clear leads under Gemini on Visual (0.7368 for S_{dens} vs. 0.6169 for RDS),

	Gemini	MiniCPM
Split	ΔAUROC (%)	ΔAUROC (%)
Audio	$+4.18^{***}$	$+1.82^{***}$
MC	-0.14 (n.s.)	$+0.85^{***}$
MO	$+4.04^{***}$	-1.20^{**}
Visual	$+0.53^{***}$	$+1.19^{***}$
Avg	$+2.15$	$+0.66$

Table 3: Mean paired relative difference in AUROC (%) of $\mathcal{C}_{\text{chain}}$ against self-consistency across 100 randomised runs. Positive ΔAUROC favours our method. Asterisks denote paired t-tests: $^{*}p < 0.01$, $^{**}p < 0.001$, $^{***}p < 0.0001$. Entries marked "n.s." are not significant.

MC (0.5622 for $\mathcal{C}_{\text{chain}}$ vs. 0.5434) and MO (0.6880 vs. 0.5918), a gap that widens to chance level under MiniCPM (0.6978 for $G_{\max}$ vs. 0.5135 on MO). Between-sample dispersion excels where answer diversity is informative, while the geometry of the reasoning chain excels where the difficulty lies in assessing an answer against multimodal evidence; a hybrid aggregate combining both families removes every below-chance split, at some cost in variance and calibration (Appendix N).

Third, the comparison with the self-verbalisation baselines favours our scores. With Gemini, every embedding score improves on the best verbalised AUROC on average (0.5918–0.6102 vs. 0.5803 for Self-Probing) and on the Visual and open-answer splits. The two families also differ in cost: verbalised confidence requires re-invoking the evaluated model on its own output, whereas our scores require only the external encoder (Appendix A quantifies the marginal costs).

Fourth, the omni-modal open-answer split confirms the structural weakness of the baselines that our method targets. All baselines fall below 0.64 AUROC on MO under Gemini, whereas the embedding scores reach 0.6880 (S_{smooth}, Gemini). UMPIRE's MO weakness persists in its native MiniCPM configuration (0.5875), so the difficulty of handling more than one extra-textual modality at a time appears structural rather than an artifact of the proxy-model setting.

Finally, the individual scores are heterogeneous across splits and generators. Extending the feature pool with the answer-level grounding of Equation 16 ($\mathcal{C}^{+}_{\text{chain}}$, Appendix G) raises the split-averaged gain over Self-Consistency with Gemini from $+2.15\%$ to $+2.62\%$ and makes all four splits significantly positive, the MC split moving from -0.14% (n.s.) to $+1.40\%^{***}$. The aggregated $\mathcal{C}_{\text{chain}}$ rarely attains the per-split maximum but tracks the best embedding score closely, and we present it as a reasonable default whenever a modest labelled validation set is available; Appendix F dissects which features carry its performance and Appendix H tests how its weights transfer across splits and generators. Appendix K shows results obtained on answers selected with our scores, where we show that our improvements are largest at small sampling budgets (Appendix M).

6 Conclusions

Taken together, our results suggest that geometric scores over the reasoning trace, computed with external omni-modal embeddings, provide a practical confidence signal for black-box omni-modal models, showing systematic gains over self-consistency under the closed generator and mixed, split-dependent gains under the open one, while outperforming the self-verbalisation baselines on discrimination for the closed-model generator while requiring no additional call to the evaluated model. On the omni-modal open-answer split, where most baselines including natively-run UMPIRE degrade sharply, the embedding scores remain the strongest family, indicating that the difficulty of assessing answers spanning several extra-textual modalities is structural for answer-level methods and is mitigated by reasoning-trace geometry. The verbalised baselines retain an advantage on raw calibration (largely a scale artefact under deployment-realistic calibration) and, with MiniCPM-generated answers, the strongest verbalised baseline matches our aggregate on discrimination averaged across splits. One practical caveat deserves emphasis: the Audio split with Gemini-generated answers shows that any confidence signal derived from sampled agreement can invert on a particular generator–modality combination, so per-modality discrimination should be validated on a small labelled sample before any of these signals is deployed as a gating or escalation mechanism.

AI USAGE STATEMENT

During the course of this research, generative AI tools were used as assistive instruments for tasks such as improving code quality, debugging, drafting and refining text, and supporting the implementation and execution of experiments. These tools were also used to facilitate literature exploration and to improve the clarity and presentation of the manuscript.

All scientific contributions, including the formulation of the research questions, the proposed methodology, experimental design, evaluation protocol, interpretation of results, and conclusions, were conceived and made by the authors. AI-generated suggestions, code, or text were critically reviewed, modified where necessary, and verified prior to inclusion. All experiments were executed, validated, and checked by the authors to ensure the correctness and reproducibility of the reported results.

The authors take full responsibility for the content of this paper, including its technical accuracy, scientific claims, and ethical compliance.

ETHICAL STATEMENT

This work proposes a method for improving uncertainty estimation in OLLMs by leveraging external embedding representations. More reliable uncertainty estimation can help users and downstream systems better identify predictions that are likely to be unreliable, thereby improving calibration, enabling selective prediction, and supporting safer deployment in high-stakes or human-in-the-loop applications.

Our method is model-agnostic and operates on learned embedding representations without modifying the underlying OLLM. As such, it does not introduce new capabilities for generating content or accessing private information beyond those already present in the base model. However, the quality of uncertainty estimates depends on both the underlying model and the external embedding model. Biases, blind spots, or distributional limitations present in either component may propagate into the estimated uncertainty. Consequently, uncertainty scores should not be interpreted as guarantees of correctness or fairness.

Improved uncertainty estimation may also be misused. For example, confidence estimates could be employed to automate decision-making in sensitive domains without appropriate human oversight, potentially leading to over-reliance on model outputs. Additionally, well-calibrated uncertainty could be exploited to selectively deploy models only when they appear confident, masking systematic failures on underrepresented populations or out-of-distribution inputs. We therefore recommend that uncertainty estimates be used as one component of a broader risk assessment framework, alongside domain expertise, robustness evaluations, and fairness analyses.

Our experiments are conducted on publicly available datasets and pretrained models. We do not collect or annotate personal data, nor do we introduce additional datasets containing sensitive information. The proposed method incurs modest additional computational overhead for extracting external embeddings and estimating uncertainty, but it does require an auxiliary embedding model at inference time, which may increase computational cost and energy consumption relative to standard inference.

Overall, we believe that improving uncertainty estimation is a positive step toward more transparent and trustworthy multimodal AI systems. Nevertheless, uncertainty estimation alone cannot address issues such as hallucination, bias, factual inaccuracies, or harmful model behavior. Future work should investigate uncertainty estimation jointly with fairness, robustness, and interpretability to better support the responsible deployment of omni-modal large language models.

REPRODUCIBILITY STATEMENT

To make our work as reproducible as possible, we have released our code via an anonymised repository and detailed our software environment, both available in Appendix O.

Further experimental details about the data used, the pre-processing steps and prompts used, when relevant, are also available in Appendix A.

Finally, more details about our baselines, in order to implement those as well in the form they appear in this work are available in Appendix B.

REFERENCES

Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sébastien Fournier. Multimodal learning with uncertainty quantification based on discounted belief fusion. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3142–3150. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/bezirganyan25a.html.

Yixin Bu, Guanyun Zou, Renzhi Wang, Runze Xia, Cunjun Wang, Hongliang Dai, Xiaoqing Ma, and Piji Li. Sampling-free uncertainty quantification via hidden state dynamics in language models. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i36.40259. URL https://doi.org/10.1609/aaai.v40i36.40259.

Chen Chen, ZeYang Hu, Fengjiao Chen, Liya Ma, Jiaxing Liu, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. Uno-bench: A unified benchmark for exploring the compositional law between uni-modal and omni-modal in omni models. *arXiv preprint arXiv:2510.18915*, 2025a. URL https://arxiv.org/abs/2510.18915.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19430–19443, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.970. URL https://aclanthology.org/2026.findings-acl.970/.

Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, ZHeng ZHang, and Richang Hong. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3095–3109, Abu Dhabi, UAE, January 2025b. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.208/.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. 2026. URL https://arxiv.org/abs/2604.27393.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9367–9385, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.558. URL https://aclanthology.org/2024.findings-acl.558/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. When cohesion lies in the embedding space: Embedding-based reference-free metrics for topic segmentation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.),

Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 17525–17536, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1524/.

Gregory Kang Ruey Lau, Hieu Dao, Nicole Kan Hui Lin, and Bryan Kian Hsiang Low. Uncertainty quantification for multimodal large language models with incoherence-adjusted semantic volume. *arXiv preprint arXiv:2602.24195*, 2026. URL https://arxiv.org/abs/2602.24195.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Min Zhang. Can LLMs learn uncertainty on their own? expressing uncertainty effectively in a self-training manner. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21635–21645, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1205. URL https://aclanthology.org/2024.emnlp-main.1205/.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Sho, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know? In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3408–3458, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.178. URL https://aclanthology.org/2026.findings-eacl.178/.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Autumn Toney and Ryan Wails. Certain but not probable? differentiating certainty from probability in LLM token outputs for probabilistic scenarios. In Bryan Eikema, Raúl Vázquez, Jonathan Berant, Marie-Catherine de Marneffe, Barbara Plank, Artem Shelmanov, Swabha Swayamdipta, Jörg Tiedemann, Chrysoula Zerva, and Wilker Aziz (eds.), *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pp. 51–60, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-349-4. doi: 10.18653/v1/2025.uncertainlp-main.6. URL https://aclanthology.org/2025.uncertainlp-main.6/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. doi: 10.48550/arXiv.2203.11171. URL https://arxiv.org/abs/2203.11171.

Chenghao Xiao, Hou Pong (Ken) Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 158370–158401. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/e839f274b3d3d682a0a82e5129d2cc4a-Paper-Conference.pdf.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

```
Given the following dictionary of exactly {num_answers} answers
extracted from a language model (keyed by their original index,
from "0" to "{num_answers - 1}"), group them based on their
semantic equivalence or having selected the same multiple-choice
option.
Return ONLY a valid JSON object with a single key 'groups'. The
value of 'groups' must be a JSON object (dictionary) mapping
EVERY index string (from "0" to "{num_answers - 1}") to an
integer group ID. You must include all {num_answers} indices,
with no omissions or extras.
Example Input Answers: {"0": "Final Answer: A", "1": "The
answer is A", "2": "Option B", "3": "B", "4": "A"}
Example Output: {"groups": {"0": 0, "1": 0, "2": 1, "3":
1, "4": 0}}
Answers to cluster (Total items: {num_answers}):
```

Figure 3: Prompt used for semantic clustering of sampled answers.

A Experimental Details

UNO-Bench. We use the UNO-Bench dataset (Chen et al., 2025a) from Hugging Face (`https://huggingface.co/datasets/meituan-longcat/UNO-Bench`) as the evaluation benchmark. The dataset contains multiple splits covering different modalities and question types, and we follow the original paper's split definitions and evaluation protocol (Chen et al., 2025a). For each question, correctness labels are computed according to the procedure described in the original UNO-Bench paper, which determines whether the generated answer matches the reference answer under the corresponding split-specific evaluation rules. The evaluated splits contain 1,000 (Audio), 1,480 (Visual), 1,000 (MC), and 250 (MO) questions, for a total of 3,730.

Answer Generation. All answers are generated using Gemini 3.1 Flash Lite (3 March 2026 version) through the `google-genai` Python library (v1.66.0) with default generation parameters. The same model configuration is used consistently for all experiments and baseline comparisons.

Omni-Modal Encoder As mentioned in the main text, we have used E5-Omni (Chen et al., 2026) as encoder for extracting external embeddings for all given input modalities. Specifically, we have used the 3B version (`https://huggingface.co/Haon-Chen/e5-omni-3B`), which underscores how relatively small encoders are already powerful enough to reach the described results.

Self-Consistency Clustering. We follow the standard self-consistency formulation introduced in Semantic Entropy (Farquhar et al., 2024), where multiple sampled generations are grouped into answer clusters to estimate agreement between model outputs. This clustering step is particularly straightforward for the Audio, Visual, and MC splits, where questions admit discrete or multiple-choice answers, while it is potentially more challenging for the MO split due to its open-ended nature.

To obtain answer clusters, we use Qwen3.5-9B (`https://huggingface.co/Qwen/Qwen3.5-9B`) as a semantic clustering model. Given the set of sampled answers, the model is prompted to group semantically equivalent responses using the instructions shown in Figure 3.

Confidence Classifier. The unified confidence score $\mathcal{C}_{\text{chain}}$ is obtained using a logistic regression classifier implemented with `scikit-learn` (Pedregosa et al., 2011). The classifier uses the extracted chain-level features as input and is trained to predict answer correctness, where correct and incorrect answers correspond to labels 1 and 0, respectively. The features are standardised per split and the model is fitted with an ℓ_1 penalty (`saga` solver, `max_iter` 5000); the inverse regularisation strength is selected by grid search over $C \in \{0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$, scoring each candidate by validation AUROC computed through the same cluster-aggregation path used at test time. The fitted model is then compared against each of the five single-feature scorers on the same validation

items and is replaced by the best single feature whenever that feature is strictly better; it is this explicit selection step, rather than the ℓ_1 penalty alone, that produces the single-feature unit-weight solutions reported in Appendix F and Appendix H. For each evaluated split, 20% of the available data is reserved as a validation set for parameter selection, while all reported test results are computed on a separate held-out portion. This separation is maintained for both our proposed methods and all baseline comparisons. Concretely, the validation sets contain 200 (Audio), 296 (Visual), 200 (MC), and 50 (MO) questions, 746 overall, leaving test sets of 800, 1,184, 800, and 200 questions, respectively (2,984 overall). The MO split is thus by far the smallest, and its per-split estimates carry correspondingly wide item-level confidence intervals (Appendix C).

Score Preprocessing. As the compared confidence measures live on heterogeneous scales (e.g., raw similarities, verbalized probabilities, entropy-derived scores), each method's confidence vector is min–max rescaled to $[0, 1]$ over the evaluated split before computing the metrics, $\tilde{c} = (c - c_{\min})/(c_{\max} - c_{\min})$. The transformation is strictly monotonic and therefore leaves AUROC and AURAC unchanged, while placing all methods on a common scale for the calibration analysis (ECE). The same preprocessing is applied uniformly to our methods and to all baselines. We note that, since the rescaling constants are computed on the evaluated split itself, the reported ECE reflects calibration after an idealized range normalization rather than the raw calibration observable at deployment; Appendix C therefore also reports a deployment-realistic variant in which a Platt calibrator is fitted on the validation items only.

Cost of the Confidence Signal. All sampling-based methods, ours included, consume the same $n = 6$ generations per question; in our data a single Gemini generation averages 2,590 tokens (median 2,146; prompt with multimodal inputs plus CoT completion), so the shared budget is roughly 15,500 evaluated-model tokens per question at metered API rates. The methods differ in the *marginal* cost of the confidence signal on top of this shared budget: zero for Self-Consistency; one forward pass of the 3B external encoder over the inputs and the sampled chains for our scores ($\approx$5 reasoning steps and an answer per chain, $\approx$240 text tokens each, plus the audio/visual inputs embedded once), which runs on a single commodity GPU and involves no further evaluated-model tokens; one additional call to the evaluated model for the self-verbalisation baselines (the question, answer, and reasoning are re-submitted, so the extra call is of the same order as a generation, $\approx$2,600 tokens); and a GPU pass of the proxy model to extract internal states for UMPIRE.

Computational Complexity. For a chain of m steps, the grounding scores require $m \times |V|$ similarities against the per-modality input embeddings and the smoothness and convergence scores $O(m)$ similarities, while the semantic-density score is the only quadratic component, at $O(m^2)$ pairwise similarities; with the typical $m \approx 5$ steps per chain observed in our data, the entire score computation is negligible next to the single forward pass of the 3B encoder over the inputs, steps, and answer.

Code and Artifact Availability. We release the full evaluation stack: feature extraction, aggregator training, the 100-run evaluation protocol, and the bootstrap significance testing, together with the extracted per-question feature vectors, the generated answers and CoT traces, and their correctness labels. The released traces are essential for exact reproduction, since the closed-source generator is served through a mutable API and its checkpoints cannot be pinned.

CoT Generation and Step Segmentation Prompts. When prompting the LLMs to generate candidate answers and then to divide the CoT into steps, we have used the prompts as in Figures 4 and 5, respectively. Step segmentation is performed by Gemini 3.1 Flash for both generators, so the two configurations are segmented identically; it is a single auxiliary pass over the already-sampled generations and involves no further call to the evaluated model.

B Baseline Implementation

UMPIRE. UMPIRE requires access to the weights of an open-source model to extract internal representations used for confidence-weighted majority voting. We use MiniCPM-O v4.5 (Cui et al., 2026) and compute the internal states of the Gemini-generated answers described in the main ex-

FINAL OUTPUT INSTRUCTIONS: You will use your internal reasoning process to analyze the inputs first. When generating your final response, you must adhere strictly to the following formatting rules:
For multiple-choice questions: Your final output must contain **ONLY** the exact letter or number of the correct choice (e.g., A, B, C, 1, 2). Do not include any rationale, restatement of the answer, markdown formatting, or disclaimers in the final output.
For open-ended questions: Provide only the direct, concise answer. Do not include introductory phrases, rationale, or disclaimers.

Figure 4: Prompt used for generating the CoT-style answers to the (pre-appended) question as per the UNO-Bench datasets.

You are given the reasoning process and final answer of a model.
REASONING: thought_text
FINAL ANSWER: answer_text
Your tasks:
Parse the reasoning above into clear, logical steps.
Return a JSON object with:
"steps": a list of strings, each representing one logical step from the reasoning

Figure 5: Prompt used for dividing the input CoT-style answer into separate reasoning steps.

periments using the `transformers` library (v5.5.4) with standard parameters. We use the length-normalised variant of the UMPIRE score, converted to a confidence as $1 - \text{UMPIRE}_{\text{len-norm}}$ (undefined values, which occur when the internal-state extraction fails for a candidate, are mapped to confidence 0). In the main Gemini experiments MiniCPM-O thus acts as a proxy model for the closed generator, while in the MiniCPM experiments (Table 24) UMPIRE runs natively on the generating model's own states.

Level-Based Confidence. For the level-based confidence baseline, we use the same Gemini model employed for answer generation (Gemini 3.1 Flash Lite, 3 March 2026 version) through the `google-genai` Python library (v1.66.0) with default parameters. The model is prompted to assign one of five qualitative confidence levels to its answer, which are mapped to the equally spaced numerical scores $\{0, 0.25, 0.5, 0.75, 1\}$ (in our data the second level was never emitted, so four distinct values occur).

Self-Probing. For Self-Probing (Liu et al., 2024), we use Gemini 3.1 Flash Lite with the same API configuration and default generation parameters. The model receives the question-answer pair and estimates the probability that the answer is correct, following the original prompting procedure.

Verbalized 2-Step CoT. For Verbalized 2-Step CoT (Tian et al., 2023), we again use Gemini 3.1 Flash Lite with the same interface and default parameters. The model is prompted to directly output the probability of correctness based on the previously generated reasoning context, without additional explanation.

C Variance, Confidence Intervals, and Deployment-Realistic Calibration

The margins separating the strongest methods in the main tables are small (of the order of 0.01 split-averaged AUROC), so this appendix makes the uncertainty around them explicit. Two distinct sources of variation are relevant, and we report both. First, *evaluation randomness*: the 100 randomised runs of the main protocol vary the train/validation/test partition and the answer-selection tie-breaking, and we report the resulting mean $\pm$ standard deviation. Second, *item-level uncertainty*: on the canonical partition (seed 42) we run a paired bootstrap over test questions ($B = 10^4$) and report percentile 95% confidence intervals. The two views are complementary and are centred differently (the interval is centred on the canonical-partition estimate, which can differ from the 100-run mean by an amount comparable to the run-to-run standard deviation), and the item-level intervals are wider by an order of magnitude: run-to-run variability is small, and generalisation to

	Gemini				MiniCPM			
	AUROC		AURAC		AUROC		AURAC	
Method	mean ± std	95% CI	mean ± std	95% CI	mean ± std	95% CI	mean ± std	95% CI
Self-Consistency	0.5991 ± 0.0047	[0.5760, 0.6178]	0.8059 ± 0.0053	[0.7841, 0.8208]	0.6380 ± 0.0041	[0.6214, 0.6600]	0.6900 ± 0.0141	[0.6773, 0.7210]
UMPIRE	0.5958 ± 0.0049	[0.5826, 0.6271]	0.7982 ± 0.0038	[0.7846, 0.8209]	0.6827 ± 0.0046	[0.6704, 0.7082]	0.7395 ± 0.0047	[0.7250, 0.7625]
Level-Based Conf.	0.5275 ± 0.0026	[0.5180, 0.5389]	0.7210 ± 0.0066	[0.7247, 0.7691]	0.6425 ± 0.0037	[0.6199, 0.6541]	0.6263 ± 0.0155	[0.6525, 0.6997]
Self-Probing Conf.	0.5838 ± 0.0042	[0.5682, 0.6009]	0.7485 ± 0.0064	[0.7541, 0.7962]	0.6978 ± 0.0047	[0.6746, 0.7121]	0.7354 ± 0.0061	[0.7111, 0.7525]
Two-Step CoT Conf.	0.5452 ± 0.0034	[0.5363, 0.5634]	0.7290 ± 0.0065	[0.7372, 0.7802]	0.6679 ± 0.0043	[0.6423, 0.6805]	0.7081 ± 0.0065	[0.6768, 0.7228]
S_{smooth}	0.6096 ± 0.0042	[0.5838, 0.6263]	0.8127 ± 0.0042	[0.7939, 0.8278]	0.6516 ± 0.0039	[0.6335, 0.6727]	0.7279 ± 0.0038	[0.7100, 0.7490]
S_{dens}	0.5988 ± 0.0043	[0.5705, 0.6137]	0.8097 ± 0.0043	[0.7907, 0.8258]	0.6549 ± 0.0040	[0.6368, 0.6757]	0.7272 ± 0.0040	[0.7089, 0.7482]
S_{goal}	0.6009 ± 0.0041	[0.5747, 0.6172]	0.8147 ± 0.0042	[0.7946, 0.8285]	0.6489 ± 0.0039	[0.6306, 0.6699]	0.7192 ± 0.0042	[0.7008, 0.7414]
G_{avg}	0.6123 ± 0.0042	[0.5869, 0.6292]	0.8139 ± 0.0042	[0.7941, 0.8278]	0.6432 ± 0.0040	[0.6250, 0.6644]	0.7227 ± 0.0039	[0.7013, 0.7418]
G_{max}	0.5935 ± 0.0048	[0.5689, 0.6112]	0.8047 ± 0.0041	[0.7846, 0.8204]	0.6369 ± 0.0041	[0.6191, 0.6596]	0.7012 ± 0.0043	[0.6774, 0.7223]
$\mathcal{C}_{\text{chain}}$	0.6081 ± 0.0110	[0.5919, 0.6340]	0.8092 ± 0.0116	[0.7952, 0.8297]	0.6446 ± 0.0132	[0.6212, 0.6619]	0.7048 ± 0.0299	[0.6669, 0.7129]

Table 4: AUROC and AURAC pooled over the union of splits (subject to the caveat of Appendix D) with both uncertainty views: mean ± sample standard deviation across the 100 randomised evaluation runs, and 95% percentile confidence intervals from a paired item-level bootstrap ($B = 10^4$) on the canonical partition. The intervals are centred on the canonical-partition estimate rather than the 100-run mean.

new questions is the binding source of uncertainty. Table 4 reports both quantities for the pooled union of the splits, the only setting for which the bootstrap was run; per-split standard deviations and intervals are included in the released result files. Pooled point estimates carry the heterogeneity caveat of Appendix D, but the run-to-run and item-level *variability* they quantify is representative of the per-split analyses as well.

Because the point estimates alone cannot settle comparisons at this scale, the tests reported in Section 5.1 operate on the paired per-question deltas rather than the marginal intervals: two methods evaluated on the same questions have strongly correlated errors, so a paired delta can be significant even when the marginal intervals overlap, as is the case for the comparison of $\mathcal{C}_{\text{chain}}$ against Self-Consistency.

Deployment-realistic calibration. The ECE values in the main tables are computed after a per-split min–max rescaling, which is rank-preserving but reflects a normalisation a practitioner could not compute at deployment. As a deployment-realistic alternative, we fit Platt scaling (a one-dimensional logistic calibrator) for every method on the *validation* items only, per split, and evaluate ECE on the calibrated test scores. Since Platt scaling is monotone, AUROC and AURAC are unaffected; only calibration changes. Table 5 reports results pooled over the union of splits (the calibrators themselves are fitted per split).

Three points follow. First, post-hoc calibration collapses ECE for every method, from the 0.15–0.63 range down to 0.007–0.07: deployment-realistic calibration is cheaply achievable across the board, and most of the large ECE differences in the main tables are scale artefacts of the raw scores rather than genuine calibration properties. Second, and in fairness to the strongest baseline, UMPIRE's severe miscalibration in the Gemini setting (0.6319) disappears entirely after calibration (0.0129); its raw ECE is an orientation artefact and should not be read as a substantive weakness. Third, the verbalised baselines remain the best-calibrated methods on Gemini even after calibration (0.007–0.009 against roughly 0.06 for the embedding-based and frequency-based scores, whose confidences take few distinct values per question and therefore cannot be calibrated within-value), so the calibration advantage we grant them in the main text survives the protocol change. We caution that a low calibrated ECE does not imply discrimination: a near-constant calibrator predicting the base rate obtains excellent ECE with chance-level AUROC, so calibrated ECE must be read jointly with the discrimination metrics.

Validation-fitted Platt slopes as an orientation diagnostic. The slope of the validation-fitted Platt calibrator provides a direct sign test of each confidence signal: a positive slope means higher confidence predicts correctness on the validation items, a negative slope means the signal is inverted. Table 6 reports the slopes on the Audio split, where the main tables show the near- or below-chance behaviour with Gemini-generated answers. With Gemini answers, the slope is negative not only for most chain-level scores but also for the encoder-independent Self-Consistency (-0.67) and for Two-Step CoT: on this generator–modality combination, agreement between sampled answers and geometric regularity of the chain are *anti-correlated* with correctness, which locates the failure in Gemini's answer distribution on audio questions rather than in the encoder or the chain-level

Method	Gemini		MiniCPM	
	min–max	calibrated	min–max	calibrated
Self-Consistency	0.2924	0.0592	0.2020	0.0573
UMPIRE	0.6319	0.0129	0.0553	0.0428
Level-Based Conf.	0.2506	0.0090	0.2573	0.0257
Self-Probing Conf.	0.2515	**0.0070**	0.2967	0.0684
Two-Step CoT Conf.	0.2533	0.0093	0.3118	0.0336
S_{smooth}	0.2888	0.0598	0.2021	0.0590
S_{dens}	0.2933	0.0647	0.2003	0.0610
S_{goal}	0.2950	0.0604	0.2003	0.0520
G_{avg}	0.2894	0.0599	0.2028	0.0541
$G_{\max}$	0.2899	0.0583	0.1541	**0.0163**
$\mathcal{C}_{\text{chain}}$	0.2869	0.0720	0.2033	0.0556

Table 5: ECE (↓) pooled over the union of splits on the canonical partition: the split-rescaled protocol of the main tables (min–max) against ECE after Platt scaling fitted on the validation split only (calibrated). Bold marks the best calibrated value per generator.

Method	Gemini		MiniCPM
	E5	LCO	E5
Self-Consistency	-0.67	-0.67	$+3.18$
UMPIRE	-0.23	-0.23	$+0.09$
Level-Based Conf.	$+2.43$	$+2.43$	$+1.82$
Self-Probing Conf.	$+10.35$	$+10.35$	$+6.65$
Two-Step CoT Conf.	-1.09	-1.09	$+3.98$
S_{smooth}	-0.08	-0.62	$+3.44$
S_{dens}	-0.25	-0.30	$+3.30$
S_{goal}	-0.10	-0.19	$+3.54$
G_{avg}	$+0.14$	-0.22	$+3.43$
$G_{\max}$	-0.29	$+4.75$	$+0.49$
$\mathcal{C}_{\text{chain}}$	$+0.14$	-0.22	$+3.45$

Table 6: Validation-fitted Platt slopes on the Audio split (200 validation items per configuration). Negative slopes indicate an inverted confidence signal. The verbalised baselines are computed once per generator and are therefore identical across the two encoder columns.

signals as such. Consistently, the same features yield strongly positive slopes and high AUROC with MiniCPM-generated answers.

	Gemini			MiniCPM		
Method	AUROC	ECE	AURAC	AUROC	ECE	AURAC
Self-Consistency	0.5991	0.2893	0.8059	0.6380	0.2023	0.6900
UMPIRE	0.5958	0.6263	0.7982	0.6827	0.0543	**0.7395**
RDS	0.6360	0.2351	0.8176	0.6420	**0.0454**	0.7195
Semantic Volume	**0.6365**	**0.1696**	**0.8187**	0.6542	0.1849	0.7296
Level-Based Conf.	0.5275	0.2497	0.7210	0.6425	0.2550	0.6263
Self-Probing Conf.	0.5838	0.2513	0.7485	**0.6978**	0.2921	0.7354
Two-Step CoT Conf.	0.5452	0.2516	0.7290	0.6679	0.3098	0.7081
S_{smooth}	0.6096	0.2846	0.8127	0.6516	0.2008	0.7279
S_{dens}	0.5988	0.2879	0.8097	0.6549	0.1987	0.7272
S_{goal}	0.6009	0.2907	0.8147	0.6489	0.1986	0.7192
G_{avg}	0.6123	0.2853	0.8139	0.6432	0.2038	0.7227
$G_{\max}$	0.5935	0.2863	0.8047	0.6369	0.2042	0.7012
$\mathcal{C}_{\text{chain}}$	0.6081	0.2835	0.8092	0.6446	0.1996	0.7048

Table 7: Pooled (Overall) AUROC (↑), ECE (↓), and AURAC (↑) on the union of the four UNO-Bench splits, under the main protocol of Tables 1 and 2 (means over 100 randomised runs). Bold marks the best value per column. These numbers are reported for completeness only: as discussed in this appendix, pooled metrics over so heterogeneous a pool are biased summaries, and the per-split values and their average in the main tables should be preferred.

Generator	Protocol	ΔAUROC (%)	ΔECE (%)	ΔAURAC (%)
Gemini	main	$+1.51^{***}$	-2.00^{***}	$+0.41^{*}$
Gemini	weighted	$+0.96^{***}$	-1.48^{***}	$+0.33$ (p=0.014)
MiniCPM	main	$+1.02^{***}$	-1.33^{***}	$+2.18^{***}$
MiniCPM	weighted	$+1.71^{***}$	-1.30^{***}	$+3.35^{***}$

Table 8: Pooled (Overall) mean paired relative differences (%) of $\mathcal{C}_{\text{chain}}$ with respect to Self-Consistency across the 100 randomised runs, for the main (grading) protocol and the geometry-weighted voting protocol of Appendix K. Asterisks follow the convention of Table 3. Being paired differences between methods evaluated on the same pool, these deltas partially cancel the shared between-split component discussed below, but they remain pooled statistics.

D Pooled (Overall) Results, and Why Pooled AUROC Is Not a Reliable Summary

Earlier versions of this work summarised performance with an "Overall" column, computed by pooling the four UNO-Bench splits into a single evaluation set of 3,730 questions (2,984 test items after the validation cut) and computing AUROC, ECE, and AURAC on the union. This appendix reports those pooled results for completeness (Table 7 for the main protocol, Table 8 for the paired deltas against Self-Consistency) and explains why we no longer use them as a headline statistic, replacing them throughout the paper with the unweighted per-split average (Avg).

Why pooled AUROC is biased on heterogeneous data. AUROC is the probability that a randomly drawn correctly-answered question receives a higher confidence than a randomly drawn incorrectly-answered one. When the evaluation set is the union of splits that differ sharply in difficulty and answer format — as UNO-Bench's do, with accuracies of the scored answers ranging from roughly 0.4 on the open-answer omni-modal split to over 0.8 on Visual — the majority of these correct–incorrect pairs are *cross-split* pairs. Such pairs are ranked correctly by any score that merely assigns systematically higher values on the easier splits, regardless of whether it can order correct above incorrect *within* any split. Pooled AUROC therefore decomposes, informally, into a within-

	Gemini					MiniCPM				
Method	Avg	Audio	MC	MO	Visual	Avg	Audio	MC	MO	Visual
Self-Consistency	0.7282	0.8188	0.6842	0.5236	0.8860	0.6865	0.8500	0.4135	0.7149	0.7677
UMPIRE	0.7240	**0.8940**	**0.6920**	0.5019	0.8080	0.6968	**0.8816**	0.4451	0.6621	**0.7985**
Level-Based Conf.	0.6760	0.8353	0.6388	0.4156	0.8144	0.6573	0.7918	0.4343	0.6652	0.7380
Self-Probing Conf.	0.7042	0.8574	0.6704	0.4537	0.8352	0.6931	0.8311	**0.4815**	0.6733	0.7866
Two-Step CoT Conf.	0.6885	0.8486	0.6503	0.4354	0.8198	0.6695	0.8234	0.4692	0.6402	0.7453
S_{smooth}	0.7276	0.8112	0.6843	0.5199	0.8951	0.6976	0.8630	0.4273	0.7310	0.7691
S_{dens}	0.7268	0.8120	0.6752	**0.5254**	0.8945	0.6924	0.8597	0.4219	0.7191	0.7690
S_{goal}	**0.7315**	0.8165	0.6901	0.5206	**0.8987**	0.6866	0.8463	0.4027	0.7265	0.7709
G_{avg}	0.7292	0.8120	0.6874	0.5206	0.8969	**0.7015**	0.8608	0.4314	**0.7339**	0.7798
$G_{\max}$	0.7268	0.8107	0.6900	0.5179	0.8888	0.6913	0.8445	0.4265	0.7223	0.7719
$\mathcal{C}_{\text{chain}}$	0.7275	0.8125	0.6864	0.5198	0.8914	0.6944	0.8613	0.4240	0.7211	0.7714

Table 9: AURAC (↑) across the UNO-Bench splits for both answer generators (same protocol as Table 1). Avg is the unweighted mean of the four per-split values (pooled-union results in Appendix D). Bold indicates the best result per column.

split component (the discrimination ability we actually care about) and a between-split component that only measures how well the score's scale happens to track split difficulty. It is a well-known failure mode of AUROC on heterogeneous pools, closely related to Simpson's paradox: a method can post a high pooled AUROC by "ranking the easiest examples first" while possessing no deeper confidence-modelling capability, and pooled rankings can even invert the per-split ordering of two methods. The unweighted per-split average used in the main text removes the between-split component by construction, which is why we adopt it as the primary summary; pooled ECE mixes per-split miscalibrations in an analogous scale-dependent way and is treated with the same caution.

Evidence in our own results. The bias is not hypothetical; several of our pooled numbers are visibly inflated by it. First, the dispersion baselines: Semantic Volume holds the best pooled Gemini AUROC in Table 7 (0.6365, against 0.6081 for $\mathcal{C}_{\text{chain}}$), yet on the per-split average it trails the best embedding scores (0.6017 vs. 0.6102 for G_{avg}): the dispersion of answer embeddings has a split-dependent scale that happens to align with split difficulty, and much of the pooled lead is between-split ranking rather than within-split skill. Second, the verbalised baselines with MiniCPM: Self-Probing's pooled AUROC of 0.6978 shrinks to 0.6563 on the split average, below $\mathcal{C}_{\text{chain}}$ (0.6603) — verbalised confidences are naturally higher on splits where the model succeeds more often, a between-split calibration property that pooled AUROC rewards as if it were discrimination. Third, entire conclusions can flip: in the feature ablations of Appendix F the largest pooled loss comes from removing $G_{\max}$ and grounding-only beats coherence-only, whereas on the split average the largest loss comes from removing S_{smooth} and the ordering of the two families reverses; likewise, the attention-style grounding score of Appendix J appears as the best MiniCPM grounding score under pooling but trails G_{avg} on the average.

E Selective Answering: AURAC Results

Following Lau et al. (2026), we consider the scenario in which a model decides whether to progressively reject its most uncertain outputs based on the relative confidence and escalate them to a higher-level verifier. We evaluate this scenario with the Area Under the Rejection–Accuracy Curve (AURAC) (Farquhar et al., 2024): the accuracy on the retained questions as the least-confident fraction of answers is progressively rejected, averaged over rejection rates (higher is better). AURAC is rank-based, so it is unaffected by the min–max rescaling of the main protocol and by the Platt scaling of Appendix C. Table 9 reports AURAC for both answer generators under the identical protocol as Tables 1 and 2 (means over 100 randomised runs; every method scores the same majority-selected answer), and Table 10 the corresponding paired differences of $\mathcal{C}_{\text{chain}}$ against Self-Consistency.

The AURAC picture mirrors the discrimination results of the main text. With Gemini-generated answers, the chain scores are the strongest family: S_{goal} attains the best average AURAC in Table 9 (0.7315) together with the best Visual value (0.8987), while $\mathcal{C}_{\text{chain}}$'s run-level deltas against Self-Consistency are mixed across splits (Table 10): significant gains on Visual and MC are offset

Split	Gemini	MiniCPM
Audio	-0.77^{***}	$+1.33^{***}$
MC	$+0.32$ (p=0.015)	$+2.59^{***}$
MO	-0.70^{***}	$+0.89^{***}$
Visual	$+0.62^{***}$	$+0.49^{*}$
Average	-0.13	$+1.33$

Table 10: Mean paired relative ΔAURAC (%) of $\mathcal{C}_{\text{chain}}$ with respect to Self-Consistency across the 100 randomised runs, for both answer generators. Positive values favour our method. Asterisks: paired t-test, $^{*}p < 0.01$, $^{**}p < 0.001$, $^{***}p < 0.0001$. The Average row is the unweighted mean of the four per-split deltas and carries no significance test; pooled-union deltas are reported in Appendix D.

by significant losses on Audio and MO, for an average of -0.13%; among the dispersion baselines of Appendix N, Semantic Volume repairs the Audio inversion (Audio AURAC up to 0.9017), consistent with its AUROC behaviour. With MiniCPM-generated answers the run-level gains of $\mathcal{C}_{\text{chain}}$ over Self-Consistency are positive and significant on all four splits (average $+1.33\%$), and the best average AURAC in Table 9 is now attained by a chain score, G_{avg} (0.7015, against 0.6968 for native UMPIRE and 0.6944 for $\mathcal{C}_{\text{chain}}$). UMPIRE's advantage is confined to the uni-modal splits (Audio 0.8816, Visual 0.7985); on the omni-modal splits the chain scores stay clearly ahead (MO: 0.7339 for G_{avg} and 0.7310 for S_{smooth}, against 0.6621).

At the item level (paired bootstrap on the canonical partition, $B = 10^4$, as in Section 5.1; these comparisons are computed on the pooled union of splits and carry the caveat of Appendix D), $\mathcal{C}_{\text{chain}}$ improves over Self-Consistency on pooled AURAC with Gemini ($+0.010$, $p = 0.038$) and with MiniCPM ($+0.023$, $p = 0.0007$), and improves over Self-Probing, the strongest verbalised baseline, on pooled AURAC with Gemini ($+0.037$, $p = 0.0003$).

F Feature Ablations for the Aggregated Score

To identify which geometric signals carry the discriminative load of $\mathcal{C}_{\text{chain}}$, we refit the aggregator on the canonical partition (Gemini-generated answers, E5-Omni features) under leave-one-feature-out and family-only configurations, keeping the fitting protocol of the main experiments unchanged (validation items only, same regularisation and grid). This analysis is run in the main-text protocol, where the aggregator scores the fixed majority answer. The full-feature configuration reproduces the stored canonical-partition results before any ablation is evaluated; because these are canonical-partition values, they differ slightly from the 100-run means of the main tables. Table 11 reports test AUROC per split, their unweighted average (Avg), and the pooled-union AURAC (retained because per-split AURAC was not stored for the ablation runs; pooled metrics carry the caveat of Appendix D).

Three observations follow. First, the validation-selected aggregator is effectively *sparse*: per split it reduces to a single feature (G_{avg} on Audio, S_{smooth} on MO, $G_{\max}$ on Visual) or, on MC, to a constant scorer whose confidence coincides with the majority-vote share (the embedding features add nothing beyond agreement on that split), which is why many leave-one-out configurations coincide with the full model on most splits. Second, no single feature is critical for the split-averaged AUROC (the spread across configurations is 0.599–0.617) because the features partially substitute for one another; the largest individual losses come from removing S_{smooth} (-0.0134 Avg AUROC, driven by MO)

Config	AUROC per split				Avg	Pooled
	Audio	MC	MO	Visual	AUROC	AURAC
full	0.4738	0.5654	0.6981	0.7122	0.6124	0.8128
$-\,S_{\text{smooth}}$	0.4738	0.5654	0.6444	0.7122	0.5989	0.8133
$-\,S_{\text{dens}}$	0.4738	0.5654	0.6981	0.7146	0.6130	0.8157
$-\,S_{\text{goal}}$	0.4738	0.5654	0.6981	0.7122	0.6124	0.8128
$-\,G_{\text{avg}}$	0.4456	0.5654	0.6981	0.7122	0.6053	0.8129
$-\,G_{\max}$	0.4738	0.5654	0.6981	0.7292	0.6166	0.8054
coherence only	0.4456	0.5654	0.6981	0.7292	0.6096	0.8072
grounding only	0.4738	0.5654	0.6569	0.7151	0.6028	0.8162

Table 11: Feature ablations of $\mathcal{C}_{\text{chain}}$ on the canonical partition (Gemini answers, E5-Omni features): leave-one-feature-out ($-$) and single-family configurations, refitted with the published protocol. Avg AUROC is the unweighted mean of the four per-split values; the AURAC column is pooled over the union of splits (Appendix D).

and G_{avg} (-0.0071, driven by Audio), whereas removing $G_{\max}$ slightly *raises* the average ($+0.0042$, via Visual) while costing pooled AURAC (-0.0074), and coherence-only edges out grounding-only on the average (0.6096 vs. 0.6028) because its MO and Visual gains outweigh the Audio rescue that G_{avg} provides. We note that under the pooled-union protocol of Appendix D the ordering differs (there the largest loss comes from removing $G_{\max}$ and grounding-only leads), an instance of how pooling can reshuffle conclusions. Third, the two families are jointly useful: coherence carries the omni-modal open-answer split (S_{smooth}; removing it costs -0.054 MO AUROC), grounding carries Audio, and Visual is served comparably by several signals, so neither family alone reproduces the full model's profile.

Fixed hyperparameters of S_{goal}. The convergence score's weighting $\gamma = 0.6$ and its $m/2$ forward/backward split point were fixed a priori and not tuned. A direct sensitivity sweep requires the per-step embeddings, which are not retained in the stored artifacts (only the step texts are), and is therefore left to the code release; we note, however, that these hyperparameters enter the pipeline only through S_{goal}, and on the canonical partition the fitted aggregator assigns S_{goal} zero marginal contribution (the $-\,S_{\text{goal}}$ row is identical to the full model), so the aggregate results are insensitive to their exact values.

G Supervision, Feature Selection, and Modality Use

This appendix collects four analyses that quantify what the aggregated score $\mathcal{C}_{\text{chain}}$ actually requires and what the grounding score actually measures. All numbers are computed offline from the stored artifacts on the canonical partition (split seed 42, validation fraction 0.2) or, where indicated, over the same 100 randomised runs as the main tables, replaying the published protocol; every analysis is gated on reproducing a stored published quantity before any new number is reported.

The oracle envelope behind the earlier version of Table 3. Table 3 reports the paired deltas of the deployable aggregator. For completeness, Table 12 reports the per-split *oracle envelope*: for each split, the largest 100-run mean relative AUROC delta over $\{S_{\text{smooth}}, S_{\text{goal}}, S_{\text{dens}}, G_{\text{avg}}, G_{\max}, \mathcal{C}_{\text{chain}}\}$, together with the score that attains it. Because the selecting score differs from split to split and is chosen on the test split, this envelope is an upper bound that no single deployable configuration attains, and it should not be read as the performance of our method: the split-averaged envelope is $+4.29\%$ (Gemini) and $+2.61\%$ (MiniCPM) against $+2.15\%$ and $+0.66\%$ for $\mathcal{C}_{\text{chain}}$.

How much labelled data does the aggregator need? $\mathcal{C}_{\text{chain}}$ is fitted on labelled validation items, which is the one supervised ingredient of the pipeline. We therefore subsample the canonical par-

	Gemini		MiniCPM	
Split	ΔAUROC (%)	attained by	ΔAUROC (%)	attained by
Audio	$+6.78^{***}$	G_{avg}	$+2.53^{***}$	G_{avg}
MC	-0.14 (n.s.)	$\mathcal{C}_{\text{chain}}$	$+3.06^{***}$	S_{smooth}
MO	$+8.66^{***}$	S_{smooth}	$+2.88^{***}$	$G_{\max}$
Visual	$+1.85^{***}$	S_{dens}	$+1.96^{***}$	S_{dens}
Avg	$+4.29$	—	$+2.61$	—

Table 12: Per-split *oracle* envelope of the relative AUROC gain over Self-Consistency across the 100 randomised runs: each cell reports the best individual score on that split and names it. The selection is made on the test split, so this is a leakage ceiling and not a deployable method; the deployable aggregator is Table 3. Significance conventions as in Table 3.

	Calibration-set size n				
Split	25	50	100	200	all
Gemini					
Audio	0.4511 ± 0.0192	0.4498 ± 0.0198	0.4589 ± 0.0177	0.4738	0.4738
MC	0.5611 ± 0.0126	0.5608 ± 0.0087	0.5623 ± 0.0068	0.5654	0.5654
MO	0.6670 ± 0.0219	0.6981 ± 0.0000	0.6981 ± 0.0000	0.6981	0.6981
Visual	0.7207 ± 0.0072	0.7215 ± 0.0071	0.7224 ± 0.0078	0.7197 ± 0.0076	0.7122
MiniCPM					
Audio	0.7278 ± 0.0095	0.7315 ± 0.0096	0.7331 ± 0.0107	0.7402	0.7402
MC	0.5532 ± 0.0055	0.5517 ± 0.0064	0.5576 ± 0.0070	0.5622	0.5622
MO	0.6890 ± 0.0201	0.7106 ± 0.0000	0.7106 ± 0.0000	0.7106	0.7106
Visual	0.6927 ± 0.0044	0.6914 ± 0.0057	0.6935 ± 0.0054	0.6936 ± 0.0023	0.6932

Table 13: Test AUROC of $\mathcal{C}_{\text{chain}}$ as a function of the number of labelled calibration examples used to fit it (mean $\pm$ standard deviation over 20 random draws of the calibration subset; entries without a deviation are the cases in which n already exhausts the split's validation items, so a single draw exists). Canonical partition, published fitting protocol.

tition's validation items of each split to $n \in \{25, 50, 100, 200\}$ (20 random draws per n), refit with the published protocol, and evaluate on the unchanged test items. Table 13 reports the resulting test AUROC. Performance is essentially flat in n: 25 labelled examples already recover the full-validation result to within the draw-to-draw standard deviation on seven of the eight configurations, and on Gemini-Visual, MiniCPM-Audio and MiniCPM-Visual the small-n fits are in fact slightly *better* than the full-validation fit, because the selection then commits less firmly to a feature that does not transfer to the test items. The practical requirement is therefore a calibration set of a few dozen labelled examples on the target configuration, not a conventional training set.

Does the learned combination beat plain feature selection? Since the fitted aggregator collapses to a single feature in most published fits, the relevant question is whether the multivariate fit adds anything over honestly selecting one feature on the validation set. Table 14 places $\mathcal{C}_{\text{chain}}$ between two references: the best single feature chosen on validation (the honest selection baseline), and an entirely unsupervised uniform combination of the five standardised features. $\mathcal{C}_{\text{chain}}$ coincides with the validation-selected single feature on six of the eight configurations and improves on it only on the two MC splits ($+0.0014$ with Gemini, $+0.0194$ with MiniCPM); it never exceeds the test-selected oracle feature, which it trails by up to 0.0225 AUROC (MiniCPM-Audio). The unsupervised uniform combination is clearly worse almost everywhere, by up to 0.2371 AUROC (Gemini-MO). The supervision is therefore doing feature *selection* rather than weight learning, and the contribution of the framework should be read as a family of interpretable geometric signals plus a lightweight, cheap procedure for choosing the reliable one for a given model and modality, rather than as a synergistic multivariate combination.

	Individual scores					Deployable			Oracle
Split	S_{smooth}	S_{goal}	S_{dens}	G_{avg}	$G_{\max}$	best single (val.)	$\mathcal{C}_{\text{chain}}$	uniform	best single (test)
Gemini									
Audio	0.4456	0.4474	0.4244	0.4738	0.4274	0.4738	0.4738	0.4858	0.4738
MC	0.5640	0.5491	0.5381	0.5606	0.5615	0.5640	0.5654	0.5334	0.5640
MO	0.6981	0.6444	0.6569	0.6852	0.6558	0.6981	0.6981	0.4610	0.6981
Visual	0.7261	0.7300	0.7292	0.7257	0.7122	0.7122	0.7122	0.6247	0.7300
MiniCPM									
Audio	0.7279	0.7306	0.7345	0.7401	0.7166	0.7401	0.7402	0.6632	0.7401
MC	0.5622	0.5449	0.5603	0.5510	0.5444	0.5622	0.5622	0.5288	0.5622
MO	0.6839	0.6638	0.6670	0.6872	0.7106	0.7106	0.7106	0.4836	0.7106
Visual	0.6891	0.6961	0.6932	0.6845	0.6801	0.6932	0.6932	0.6437	0.6961

Table 14: Test AUROC on the canonical partition of the individual scores, of the best single score selected on the validation items, of $\mathcal{C}_{\text{chain}}$, of an unsupervised uniform combination of the five standardised features, and of the best single score selected on the test items. The last column uses test labels and is a leakage ceiling, not a method.

	MC		MO	
Grounding aggregation	AUROC	AURAC	AUROC	AURAC
Maximum over modalities (Def. 1)	0.5530	0.6741	0.6719	0.5175
Mean over modalities	0.5524	0.6743	0.6693	0.5164
Geometric mean over modalities	0.5523	0.6744	0.6698	0.5166
Variance-penalised	0.5524	0.6742	0.6699	0.5166
Entropy-weighted	0.5550	0.6746	0.6711	0.5173
Minimum over modalities	**0.5585**	0.6745	**0.6726**	0.5168
Single stream (first)	0.5618	0.6747	0.6751	0.5173
Single stream (second)	0.5517	0.6762	0.6717	0.5172

Table 15: Grounding aggregation rules on the two genuinely bi-modal splits (100-run means, published protocol, Gemini answers). The two single-stream rows are the per-modality alignments taken alone; we do not label them audio and visual because the stored ordering of the per-modality alignments could not be verified against the stored values, the between-stream differences being of the order of 0.001–0.005.

Is the grounding score using more than one modality? Definition 1 aggregates the per-modality alignments with a maximum, which rewards a step aligned with a single modality and never requires simultaneous support from several. We test the consequences directly on the splits where more than one modality is present. Of the 3,730 questions, only MC (1,000) and MO (250) carry two modalities throughout; Audio is single-modality by construction and 1,429 of the 1,480 Visual questions carry only an image or a video, so on those splits every aggregation rule coincides with the maximum by construction and the comparison is uninformative. Table 15 therefore reports the MC and MO columns, evaluated under the published protocol over the same 100 runs, using a parallel scoring run of the Gemini data that retains the per-modality alignments (it shares the step segmentation of the published run up to holding the last step out as the answer unit, so its max-aggregation is a parallel G_{avg} rather than a replay of the published value; all variants are compared within that run).

Two results follow. First, each individual modality stream is separately predictive — the single-stream scores reach 0.5517–0.5618 AUROC on MC and 0.6717–0.6751 on MO, against 0.5530 and 0.6719 for the max — so the grounding signal is not an artefact of one dominant modality. Second, requiring *simultaneous* support is mildly but significantly better than the maximum: taking the minimum over modalities, the strictest form of joint grounding, improves MC AUROC from 0.5530 to 0.5585 ($p < 10^{-4}$) and the split-pooled value from 0.6008 to 0.6044, while the softer synthesis rules (geometric mean, variance-penalised, entropy-weighted aggregation) sit within ± 0.003 of the maximum. The effect sizes are small, and we report them as evidence that the max-aggregation of Definition 1 does under-use cross-modal agreement rather than as a replacement for it.

Short chains and answer ties. Two protocol quantities referenced in Section 3 and Section 4 are reported here. Single-step chains, for which S_{smooth}, S_{dens} and Δ_{bwd} are undefined and our implementation assigns the maximal value, number 17 of 24,016 candidate chains (0.07%) with Gemini but 6,151 of 37,300 (16.5%) with MiniCPM, distributed as 265/10,000 on Audio, 2,114/10,000 on MC, 501/2,500 on MO and 3,271/14,800 on Visual. The MiniCPM coherence scores should therefore be read with the caveat that a sixth of its candidate chains receive maximal coherence by convention. Questions whose count-majority cluster is not unique, and whose returned answer is consequently decided by each method's own tie-break, number 1,079 of 3,730 (28.9%) with Gemini (601/1,000 Audio, 243/1,000 MC, 139/250 MO, 96/1,480 Visual) and 529 of 3,730 (14.2%) with MiniCPM (123/1,000 Audio, 104/1,000 MC, 140/250 MO, 162/1,480 Visual).

Answer-level grounding, formally. For a candidate generation g with embedding $\phi_g = \mathcal{E}(g)$ and the modality alignment $A^{(k)}(g) = \max_{\phi_v \in \Phi^{(k)}} \mathcal{S}(\phi_g, \phi_v)$ of Equation 2, the three answer-level grounding scores used in this appendix are

$$G_{\text{ans}}^{\max} = \max_k A^{(k)}(g), \quad G_{\text{ans}}^{\text{avg}} = \frac{1}{K}\sum_k A^{(k)}(g), \quad G_{\text{ans}}^{\min} = \min_k A^{(k)}(g). \tag{16}$$

They read the generation as a whole rather than step by step, and so discard the trace, but they are computed from the same encoder, the same input pool and the same similarity as the chain-level scores.

Extending the feature pool with answer-level grounding. The analyses above locate the aggregator's weakness in *which* signals it has to choose from rather than in how it combines them. We therefore refit it, with the published protocol unchanged, on the feature vector of Equation 12 extended by the three answer-level grounding scores of Equation 16, and evaluate over the same 100 randomised runs. Table 16 reports the result. With Gemini the split-averaged relative AUROC gain over Self-Consistency rises from $+2.15\%$ to $+2.62\%$, every split becomes significantly positive, and the multiple-choice split moves from a non-significant -0.14% to $+1.40\%$; the open-answer omni-modal split gains $+4.53\%$. With MiniCPM the average rises from $+0.66\%$ to $+0.84\%$ and AURAC improves on all four splits, and the open-answer split's AUROC loss becomes small and no longer significant (-0.79%, $p = 0.11$). The chain-only configuration refit inside this experiment reproduces the published per-split values to within 0.0007 AUROC, so the two columns are directly comparable.

Can the selection step itself be improved? The analyses above locate the aggregator's weakness in the *selection* of a signal rather than in how signals are combined, and the per-split validation sets are small (50 questions on MO). We therefore evaluated six alternatives to the argmax over validation AUROC, all of which see validation labels only, over the same 100 randomised runs on the eight-feature pool. Three attack the variance of the selection: a 5-fold cross-validated estimate of each feature's validation AUROC before the argmax (*cv*), a soft vote over the argmaxes of 30 bootstrap resamples of the validation items (*bagged*), and selection on the pooled validation items of all four splits (*pooled*). Three replace hard selection by a weighting of the per-question feature ranks: softmax weights over validation AUROC (*softmax*), the mean rank of the top-k features with k chosen on validation (*top-k*), and the uniform combination already reported above. Table 17 reports the split-averaged relative AUROC gain over Self-Consistency.

Three conclusions follow, and the third is a negative result worth stating plainly. First, *widening the pool helps more than changing the rule*: every selector that keeps a hard argmax over the eight features beats the published five-feature aggregator on both generators, and the single largest step is the addition of the answer-level grounding scores. Second, *reducing selection variance helps only marginally*: the cross-validated argmax is within 0.15 points of the plain argmax on both generators, and pooling the validation items across splits helps Gemini's smallest split but costs MiniCPM 1.3 points on average, because it forces one feature onto four splits whose best signals genuinely differ. Third, *rank-based soft combination is not a safe default*: it is the best selector on Gemini ($+3.02$) and the worst on MiniCPM (-2.04). The asymmetry has a concrete cause. The confidence attached to an answer is the sum of its cluster's per-candidate scores normalised by the total, a quantity that for narrow-band scores is dominated by the cluster's size and therefore carries the agreement signal itself. Replacing the raw values by per-question ranks removes that magnitude information;

	AUROC			AURAC		
Split	Self-Cons.	$\mathcal{C}_{\text{chain}}$	$\mathcal{C}^{+}_{\text{chain}}$	Self-Cons.	$\mathcal{C}_{\text{chain}}$	$\mathcal{C}^{+}_{\text{chain}}$
Gemini						
Audio	0.4512	0.4696	0.4700***	0.8188	0.8126	0.8126
MC	0.5630	0.5626	**0.5709*****	0.6842	0.6863	**0.6940*****
MO	0.6334	0.6584	**0.6621*****	0.5236	0.5199	0.5221
Visual	0.7232	0.7272	0.7257***	0.8858	0.8915	0.8893***
Avg	0.5927	0.6044	**0.6072**	0.7281	0.7276	0.7295
MiniCPM						
Audio	0.7236	0.7368	**0.7372*****	0.8519	**0.8612**	0.8587***
MC	0.5461	**0.5505**	0.5489*	0.4138	**0.4238**	0.4230***
MO	**0.6782**	0.6711	0.6727	0.7154	0.7217	**0.7241*****
Visual	0.6755	0.6838	**0.6874*****	0.7697	0.7717	**0.7763*****
Avg	0.6559	0.6606	**0.6616**	0.6877	0.6946	**0.6955**

Table 16: The aggregator refit on the chain features alone ($\mathcal{C}_{\text{chain}}$) and on the same features extended with the three answer-level grounding scores of Equation 16 ($\mathcal{C}^{+}_{\text{chain}}$), 100-run means under the published protocol. Asterisks mark the paired t-test of $\mathcal{C}^{+}_{\text{chain}}$ against Self-Consistency with the conventions of Table 3; bold marks the best of the three per row.

Selector	Gemini	MiniCPM
$\mathcal{C}_{\text{chain}}$, five chain features	$+2.12$	$+0.68$
$\mathcal{C}^{+}_{\text{chain}}$, eight features	$+2.60$	$+0.84$
argmax of validation AUROC	$+2.79$	$+\mathbf{0.98}$
cross-validated argmax	$+2.66$	$+0.90$
pooled-validation argmax	$+2.49$	-0.30
softmax-weighted ranks	$+\mathbf{3.02}$	-2.04
bagged selection	$+2.58$	-1.62
top-k mean rank	$+2.08$	-2.38

Table 17: Split-averaged relative AUROC gain over Self-Consistency (%) for alternative selectors over the eight-feature pool, 100 randomised runs, validation-only selection throughout.

where agreement is weak (Gemini, Self-Consistency at 0.5927 average AUROC) nothing is lost, but where it is strong (MiniCPM, 0.6559) the combination discards the very signal it should be refining. We therefore retain the published protocol, and read these results as evidence that the aggregator's headroom lies in the quality and breadth of the individual signals rather than in the sophistication of their combination.

H Transferability of the Learned Aggregation Weights

The aggregated score $\mathcal{C}_{\text{chain}}$ is fitted per split on labelled validation data, so a natural question is whether the learned weights transfer to splits or answer-generating models other than the one they were fitted on. We test both directions on the canonical partition: for every ordered pair of splits we apply the aggregator fitted on the source split's validation items to the target split's test items (cross-split), and we apply each split's Gemini-fitted aggregator to the corresponding MiniCPM test

items and vice versa, aligning features by semantic role (cross-model). The analysis is run in the main-text protocol; the replay reproduces the stored in-split values before any transfer is evaluated.

The fitted aggregator is nearly always a single-feature selector. A structural observation frames the results: in 8 of the 10 fits (4 splits plus pooled, for each generator) the validation-selected aggregator collapses to a single feature with unit weight (e.g., G_{avg} on Gemini-Audio and on MiniCPM-Audio, S_{smooth} on Gemini-MO, $G_{\max}$ on MiniCPM-MO), and on Gemini-MC the fitted logistic regression is effectively constant, reducing the confidence to the majority-vote share. Transferring $\mathcal{C}_{\text{chain}}$ therefore amounts, in practice, to transferring a validation-selected choice of the best single chain-level feature.

Results. Transfer is close to free on most pairs: the mean off-diagonal change in AUROC under cross-split transfer is -0.015 (Gemini) and -0.001 (MiniCPM), with all AURAC changes within ± 0.026, and cross-model transfer changes pooled AUROC by at most ± 0.006 (Gemini$\rightarrow$MiniCPM $+0.003$, MiniCPM$\rightarrow$Gemini -0.005). The failures concentrate exactly where the split-specific feature choice was the learned content: aggregators fitted on MC lose 0.04–0.06 AUROC when applied to MO, and MiniCPM-fitted weights lose 0.046 AUROC on Gemini-Audio, where MiniCPM's selected feature is near-worst for the other generator. Transferring fits obtained on the pooled union of splits across models is uniformly worse (-0.02 AUROC in both directions) than transferring the per-split fits. Two caveats apply: the Gemini-MC in-split value rests on tie-breaking of an effectively constant scorer and comparisons on that cell are noise; and even in-split, $\mathcal{C}_{\text{chain}}$ trails the best label-free single score on some splits, so transfer deltas of ± 0.02 are small relative to the gap to oracle per-split score selection. Full matrices are included in the released result files.

The practical reading is twofold: the learned weights are not model-specific in any deep sense, so reusing them across generators is safe *on average*; but because what they encode is a per-split feature choice, they inherit the feature-instability limitation discussed in Appendix G, and a small labelled validation sample on the target configuration remains the reliable way to select or fit the aggregate — one that Appendix G shows can be as small as a few dozen examples.

I Robustness to the Choice of Omnimodal Encoder

All embedding-based scores in the main experiments are computed from E5-Omni (Chen et al., 2026) representations. To verify that our findings do not hinge on this particular encoder, we repeat the main-text evaluation (same Gemini-generated candidates, same protocol and 100 randomised runs) replacing E5-Omni with LCO-Embedding-Omni-7B (Xiao et al., 2025), a 7-billion-parameter omnimodal embedding model trained with a language-centric representation-learning recipe that differs from E5-Omni in training procedure, training data, and scale (7B vs. 3B parameters). Methods that do not depend on the encoder (Self-Consistency, UMPIRE, and the self-verbalisation baselines) are unaffected by this substitution and their rows repeat the main results; we include them for ease of comparison. Tables 18 and 19 mirror Tables 1 and 9.

Two observations emerge. First, the headline conclusions are stable under the change of encoder. On the split average, $\mathcal{C}_{\text{chain}}$ moves from an AUROC of 0.6046 with E5-Omni to 0.6008 with LCO and from an AURAC of 0.7275 to 0.7284; it remains ahead of Self-Consistency, UMPIRE, and all verbalised baselines on both average metrics, and the embedding scores remain the strongest family on the open-answer omni-modal split, where S_{goal} with LCO attains the best MO AUROC observed with Gemini-generated answers (0.6917, against 0.6880 for the best E5-Omni score). For the aggregate, the absolute difference between encoders never exceeds 0.03 on any split–metric combination, and on Visual and MO the larger encoder is slightly beneficial.

Second, while the aggregate is stable, *which* individual score leads shifts with the encoder, consistent with the heterogeneity observed in the main text. With LCO, S_{goal} replaces G_{avg} as the best average discriminator (0.6036, with G_{avg} close behind), and S_{smooth} attains the best average, MC, and Visual AURAC. The grounding scores are the most encoder-sensitive: G_{avg} loses its average lead (0.6102 to 0.6025), whereas $G_{\max}$ gains substantially on the multiple-choice and Visual splits (0.5582 to 0.5593 and 0.7232 to 0.7432, the best Visual AUROC in the table) and attains the best MO AURAC (0.5338), while losing ground on Audio (0.4351 to 0.4372 remains below chance). That individual scores fluctuate with the encoder while $\mathcal{C}_{\text{chain}}$ does not is precisely the behaviour the aggregation was

	Avg		Audio		MC		MO		Visual	
Method	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE
Self-Consistency	0.5927	0.3068	0.4512	0.6203	**0.5630**	0.2723	0.6334	0.2459	0.7234	**0.0888**
UMPIRE	0.5825	0.4507	**0.6318**	0.7075	0.5551	0.3999	0.6159	**0.1307**	0.5272	0.5649
Level-Based Conf.	0.5230	0.3193	0.5074	0.1661	0.5154	0.3466	0.5124	0.5777	0.5566	0.1866
Self-Probing Conf.	0.5804	0.3170	0.5705	**0.1594**	0.5495	0.3378	0.5871	0.5688	0.6143	0.2019
Two-Step CoT Conf.	0.5435	0.3147	0.5441	0.1611	0.5109	0.3473	0.5411	0.5522	0.5779	0.1981
S_{smooth}	0.5977	0.3013	0.4335	0.6116	0.5571	0.2754	0.6629	0.2264	0.7373	0.0917
S_{dens}	0.5832	0.2973	0.4257	0.6016	0.5432	0.2792	0.6266	0.2149	0.7374	0.0935
S_{goal}	**0.6036**	**0.2968**	0.4479	0.6064	0.5486	0.2872	**0.6917**	0.2035	0.7260	0.0899
G_{avg}	0.6025	0.3013	0.4554	0.6219	0.5530	0.2756	0.6719	0.2153	0.7297	0.0924
$G_{\max}$	0.5993	0.3022	0.4372	0.6139	0.5593	**0.2674**	0.6574	0.2378	**0.7432**	0.0896
$\mathcal{C}_{\text{chain}}$	0.6007	0.3013	0.4409	0.6178	0.5611	0.2741	0.6654	0.2206	0.7352	0.0925

Table 18: AUROC (↑) and ECE (↓) across the UNO-Bench splits with all embedding-based scores computed from LCO-Embedding-Omni-7B instead of E5-Omni (Gemini-generated answers, same protocol as Table 1). Avg as in Table 1; bold indicates the best result per column.

Method	Avg	Audio	MC	MO	Visual
Self-Consistency	0.7282	0.8188	0.6842	0.5236	0.8860
UMPIRE	0.7240	**0.8940**	0.6920	0.5019	0.8080
Level-Based Conf.	0.6760	0.8353	0.6388	0.4156	0.8144
Self-Probing Conf.	0.7042	0.8574	0.6704	0.4537	0.8352
Two-Step CoT Conf.	0.6885	0.8486	0.6503	0.4354	0.8198
S_{smooth}	**0.7295**	0.8134	**0.6966**	0.5099	**0.8983**
S_{dens}	0.7249	0.8102	0.6890	0.5025	0.8981
S_{goal}	0.7269	0.8117	0.6849	0.5207	0.8902
G_{avg}	0.7268	0.8203	0.6741	0.5175	0.8952
$G_{\max}$	0.7277	0.8162	0.6669	**0.5338**	0.8938
$\mathcal{C}_{\text{chain}}$	0.7284	0.8165	0.6841	0.5199	0.8931

Table 19: AURAC (↑) across the UNO-Bench splits with LCO-Embedding-Omni-7B (Gemini-generated answers). Avg is the unweighted mean of the four per-split values. Bold indicates the best result per column.

designed for, and we regard this as evidence that the proposed method is robust to the choice of the external omnimodal embedding model.

J Attention-Style Grounding: Exercising the Per-Modality Alignment Profile

The remark in the Methodology notes that, under a single shared encoder, the max-pooled omni-modal grounding coincides with its uni-modal counterpart, and that the value of the partitioned formulation lies in enabling aggregations that actually use the per-modality alignment profile $\{A^{(k)}(r_i)\}_{k=1}^{K}$. This appendix evaluates one such aggregation, an attention-style grounding score G_{att}, against the simple grounding scores G_{avg} and $G_{\max}$ of the main text.

Definition. The score has two instantiations, matching the structure of the input. In the *omni-modal* setting ($K > 1$), each reasoning step is softly routed to its best modality and rewarded for being decisive about it (entropy-gated routing): with $p_i = \mathrm{softmax}_k\big(A^{(k)}(r_i)/\tau\big)$ the routing

Configuration	Score	Avg		Audio		MC		MO		Visual	
		AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
Gemini, E5-Omni	G_{avg}	**0.6102**	**0.7292**	**0.4817*****	0.8120	0.5576	0.6874	**0.6705*****	**0.5206*****	**0.7311*****	**0.8969*****
	$G_{\max}$	0.5918	0.7268	0.4351	0.8107	**0.5582**	**0.6900*****	0.6507	0.5179	0.7232	0.8888
	G_{att}	0.5675	0.7107	0.4211	**0.8145*****	0.5498	0.6674	0.5733	0.4693	0.7256	0.8916
Gemini, LCO	G_{avg}	**0.6025**	0.7268	0.4554	**0.8203**	0.5530	**0.6741**	**0.6719*****	0.5175	0.7297	**0.8952*****
	$G_{\max}$	0.5993	**0.7277**	0.4372	0.8162	**0.5593*****	0.6669	0.6574	**0.5338*****	**0.7432*****	0.8938
	G_{att}	0.5938	0.7232	**0.4593*****	0.8197	0.5538	0.6736	0.6316	0.5048	0.7305	0.8949
MiniCPM, E5-Omni	G_{avg}	**0.6588**	**0.7015**	**0.7419*****	**0.8608*****	**0.5507**	**0.4314**	0.6675	**0.7339*****	0.6750	**0.7798*****
	$G_{\max}$	0.6584	0.6913	0.7223	0.8445	0.5439	0.4265	**0.6978*****	0.7223	0.6695	0.7719
	G_{att}	0.6427	0.6830	0.7304	0.8387	0.5428	0.4175	0.6216	0.7121	**0.6760*****	0.7638

Table 20: The attention-style grounding score G_{att} against the simple grounding scores, sampling-free regime, AUROC and AURAC (↑; means over 100 randomised runs). Avg is the unweighted mean of the four per-split values. Bold marks the best of the three scores per column within each configuration; asterisks follow the 100-run paired t-test convention of Table 21 and are not attached to the derived Avg column.

distribution of step r_i over modalities and $H(p_i)$ its Shannon entropy,

$$G_{\text{att}} \;=\; \frac{1}{m}\sum_{i=1}^{m}\Big(1-\tfrac{H(p_i)}{\log K}\Big)\,\max_{k} A^{(k)}(r_i), \tag{17}$$

so that steps aligning strongly with a *single* modality contribute fully, while generic steps that weakly align with everything are suppressed. In the *uni-modal* setting, where routing over modalities is vacuous, the same sharpening is applied over steps instead (salience-weighted alignment): with $w = \mathrm{softmax}_i(A(r_i)/\tau)$,

$$G_{\text{att}} \;=\; \sum_{i=1}^{m} w_i\, A(r_i), \tag{18}$$

which concentrates the score on the few most grounded anchor steps rather than averaging over the whole chain. Both variants use temperature $\tau = 0.05$.

Protocol. G_{att} is evaluated exactly like the other individual scores in the main tables (sampling-free regime, 100 randomised runs); we verified our replay of the protocol by reproducing the stored G_{avg} results of every run to numerical precision. Table 20 compares the three grounding scores across the three configurations of this work. Bold indicates the best of the three per column; asterisks follow the 100-run paired t-test convention of Table 21, computed within each block.

Findings. The picture is configuration-dependent, and instructive on both ends. With MiniCPM-generated answers, G_{att} retains the best Visual AUROC but trails both simple scores on the split average (0.6427 against 0.6588 for G_{avg} and 0.6584 for $G_{\max}$), its weakness being concentrated on MO, which is precisely the split where $G_{\max}$ is strongest (0.6978). With Gemini-generated answers and the E5-Omni encoder, the plain mean alignment G_{avg} is clearly superior (0.6102 vs. 0.5675 on average), and with the LCO encoder $G_{\max}$ leads on the split average (0.5993 against 0.6025 for G_{avg} and 0.5938 for G_{att}). The sharpened aggregation is therefore not a uniform improvement: its benefit appears to depend on how sharply the encoder separates modality alignments at the fixed temperature, and we have not tuned τ.

One pattern, however, is consistent across all three configurations: G_{att} is the *worst* of the three scores on the open-answer omni-modal split (e.g. 0.5733 vs. 0.6705 for G_{avg} with E5-Omni), precisely the split where our scores otherwise lead all baselines. This is perhaps unsurprising in retrospect: the entropy gate rewards steps that commit to a single modality, but open-answer omni-modal questions are exactly those whose reasoning must integrate audio and visual evidence within the same step, so the decisiveness prior penalises the behaviour that characterises correct answers there. In our view this is useful evidence about the design space: the per-modality alignment profile is genuinely informative (with MiniCPM it yields the best grounding score we have on the Audio and Visual splits), but a hard one-modality-per-step prior is the wrong inductive bias for genuinely omni-modal reasoning, and modality-aware aggregations that instead reward *balanced* cross-modal alignment are the natural next candidate.

Method	Avg		Audio		MC		MO		Visual	
	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE
Self-Consistency	0.5927	0.3068	0.4512	0.6203	**0.5630**	0.2723	0.6334	0.2459	0.7234	**0.0888*****
Level-Based	0.5230	0.3192	0.5074	0.1661	0.5154	0.3466	0.5124	0.5777	0.5566	0.1866
Self-Probing	0.5803	0.3170	0.5705	**0.1594*****	0.5495	0.3378	0.5871	0.5688	0.6143	0.2019
Two-Step CoT	0.5435	0.3147	0.5441	0.1611	0.5109	0.3473	0.5411	0.5522	0.5779	0.1981
UMPIRE	0.5825	0.4507	**0.6318*****	0.7075	0.5551	0.3999	0.6159	**0.1307*****	0.5272	0.5649
S_{smooth}	0.6068	**0.2980**	0.4551	0.6124	0.5614	0.2758	**0.6795*****	0.2109	0.7311	0.0928
S_{dens}	0.5923	0.3057	0.4355	0.6174	0.5405	0.2860	0.6563	0.2271	**0.7368***	0.0923
S_{goal}	0.5990	0.3052	0.4563	0.6122	0.5465	0.2895	0.6577	0.2224	0.7356	0.0969
G_{avg}	**0.6084**	0.3005	0.4769	0.6061	0.5541	0.2826	0.6699	0.2179	0.7328	0.0954
$G_{\max}$	0.5918	0.3043	0.4351	0.6148	0.5582	**0.2718**	0.6507	0.2355	0.7232	0.0952
$\mathcal{C}_{\text{chain}}$	0.6011	0.3014	0.4621	0.6107	0.5629	0.2745	0.6521	0.2267	0.7275	0.0936

Table 21: Geometry-weighted majority voting: AUROC (↑) and ECE (↓) across the UNO-Bench splits with Gemini-generated answers. Every confidence method re-weights the votes of the $n = 6$ candidates. Avg is the unweighted mean of the four per-split values (pooled-union results in Appendix D). Bold indicates the best result per column; asterisks here denote 100-run paired t-test comparisons with the runner-up ($p < 0.01$, $p < 0.001$, and $p < 0.0001$ for one, two, and three asterisks), unlike the item-level convention of the main tables, and are not attached to the derived Avg column.

K Geometry-Weighted Self-Consistency

In the main protocol the chain-level scores *grade* a fixed answer: the returned answer is the count-majority cluster of the sampled candidates, and the confidence attached to it is the normalised score-mass (sum-share) of that cluster. This appendix evaluates the natural next step, in which the same score-mass is also allowed to *select* the answer: each sampled chain votes with a weight equal to its chain-level score, so a minority cluster of well-grounded, coherent chains can overturn the count majority. The confidence formation is identical to the main protocol (the winning cluster's sum-share); the only difference is which cluster wins.

Method. Standard self-consistency treats all sampled reasoning chains equally. Given N sampled generations $\{(R^{(j)}, a^{(j)})\}_{j=1}^{N}$, we compute for each the chain-level feature vector

$$\mathbf{x}^{(j)} = \left[S^{(j)}_{\text{smooth}}, S^{(j)}_{\text{dens}}, S^{(j)}_{\text{goal}}, G^{(j)}_{\text{avg}}, G^{(j)}_{\max}\right]$$

and a confidence score $c^{(j)}$, obtained either from the logistic-regression aggregator of the main text, $c^{(j)} = \mathcal{C}_{\text{chain}}(\mathbf{x}^{(j)})$, or directly from a single feature in the individual-score ablations. The standard self-consistency estimate, $\hat{P}(y|q,V) = \frac{1}{N}\sum_j \mathbf{1}(a^{(j)} = y)$, is then replaced by its geometry-weighted counterpart:

$$\hat{P}(y|q,V) = \frac{\sum_{j=1}^{N} c^{(j)} \mathbf{1}(a^{(j)} = y)}{\sum_{j=1}^{N} c^{(j)}}, \tag{19}$$

which assigns larger weights to answers supported by coherent and well-grounded reasoning chains; the answer returned is $\arg\max_y \hat{P}(y|q,V)$ and its confidence is $\hat{P}$ itself. The relation to the main protocol is exact: there, the same $\hat{P}(y|q,V)$ is computed but the count-majority answer is returned (with $\hat{P}$ of that answer as its confidence), so the two settings differ only in whether the geometric evidence may change the selected answer, not in how confidence is formed. The evaluation protocol is identical to the main experiments ($N = 6$, 100 randomised runs).

Results. Tables 21 and 22 report the geometry-weighted results with Gemini-generated answers, in the format of the main tables; Table 23 gives the paired relative differences of $\mathcal{C}_{\text{chain}}$ against plain Self-Consistency across the 100 matched runs. The weighted $\mathcal{C}_{\text{chain}}$ raises the split-averaged AUROC from 0.5927 to 0.6011, with the largest gains on the open-answer omni-modal split (AUROC 0.6334 to 0.6521), while the split-averaged AURAC is essentially unchanged (0.7279 vs. 0.7281), the AUROC gains on Audio and MO coming with small AURAC losses on those same splits (Table 23); the run-level analysis marks the AUROC improvements as highly systematic. The gains are

Method	Avg	Audio	MC	MO	Visual
Self-Consistency	0.7281	0.8188	0.6842	0.5236	0.8860
Level-Based	0.6760	0.8353	0.6388	0.4156	0.8144
Self-Probing	0.7042	0.8574	0.6704	0.4537	0.8352
Two-Step CoT	0.6885	0.8486	0.6503	0.4354	0.8198
UMPIRE	0.7240	**0.8940*****	**0.6920**	0.5019	0.8080
S_{smooth}	0.7276	0.8112	0.6841	0.5200	0.8951
S_{dens}	0.7268	0.8120	0.6752	**0.5254**	0.8945
S_{goal}	**0.7309**	0.8161	0.6902	0.5188	**0.8987*****
G_{avg}	0.7293	0.8131	0.6870	0.5202	0.8969
$G_{\max}$	0.7268	0.8107	0.6900	0.5179	0.8888
$\mathcal{C}_{\text{chain}}$	0.7279	0.8144	0.6867	0.5192	0.8912

Table 22: Geometry-weighted majority voting: AURAC (↑) across the UNO-Bench splits with Gemini-generated answers. Avg, bold, and asterisks as in Table 21.

Split	ΔAUROC (%)	ΔECE (%)	ΔAURAC (%)
Audio	$+2.44^{***}$	-1.55^{***}	-0.54^{***}
MC	-0.03 (n.s.)	$+0.82^{*}$	$+0.36^{*}$
MO	$+2.97^{***}$	-7.71^{***}	-0.82^{***}
Visual	$+0.56^{***}$	$+5.44^{***}$	$+0.60^{***}$
Average	$+1.49$	-0.75	-0.10

Table 23: Mean paired relative difference (%) of the geometry-weighted $\mathcal{C}_{\text{chain}}$ with respect to Self-Consistency across 100 runs (Gemini-generated answers). Positive ΔAUROC/ΔAURAC and negative ΔECE favour the weighted variant. Asterisks: paired t-test, $^{*}p < 0.01$, $^{**}p < 0.001$, $^{***}p < 0.0001$. The Average row is the unweighted mean of the four per-split deltas and carries no significance test; pooled-union deltas are reported in Appendix D.

nonetheless small in absolute terms, and most of the weighted scheme's performance is inherited from the strength of self-consistency itself.

Item-level significance. Since the 100 runs share the same underlying question pool, we complement the run-level tests with a paired bootstrap over test questions on a canonical partition ($B = 10^4$), analogous to the main-text convention; these comparisons are computed on the pooled union of splits and carry the caveat of Appendix D. At the item level, the pooled gains of the weighted $\mathcal{C}_{\text{chain}}$ over Self-Consistency are significant with MiniCPM-generated answers (AUROC $+0.014$, $p = 0.006$; AURAC $+0.023$, $p = 0.0007$) and marginal with Gemini ($+0.014$, $p = 0.017$; $+0.011$, $p = 0.026$), while no single-split difference reaches significance and the result does not replicate on the canonical partition of the alternative-encoder configuration of Appendix I. One honestly-selected per-split result stands out: with the LCO encoder, validation selects $G_{\max}$ as the weighting signal on Audio, and the choice holds up at the item level ($+0.120$ AUROC over Self-Consistency, $p = 0.003$).

	Avg		Audio		MC		MO		Visual	
Method	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE
Majority Vote (Self-Consistency)	0.6559	0.2553	0.7236	0.1310	0.5461	0.3198	0.6782	0.4051	0.6755	0.1653
Level-Based Conf.	0.6200	0.2699	0.6306	0.1796	0.6024	0.3905	0.5956	0.2745	0.6512	0.2350
Self-Probing Conf.	0.6563	0.2843	0.6941	0.1753	**0.6331*****	0.4658	0.5933	0.2521	0.7047	0.2440
Two-Step CoT Conf.	0.6294	0.3133	0.6791	0.1830	0.6252	0.4804	0.5609	0.3028	0.6525	0.2871
UMPIRE	0.6570	**0.1577**	**0.7507*****	0.1550	0.5840	0.2898	0.5875	**0.1323*****	**0.7058**	**0.0538*****
S_{smooth}	0.6564	0.2628	0.7270	0.1406	0.5628	0.3116	0.6531	0.4357	0.6829	0.1632
S_{dens}	0.6538	0.2590	0.7344	**0.1293**	0.5572	0.3141	0.6347	0.4317	0.6887	0.1610
S_{goal}	0.6522	0.2627	0.7295	0.1354	0.5451	0.3155	0.6462	0.4408	0.6878	0.1593
G_{avg}	0.6588	0.2649	0.7419	0.1346	0.5507	0.3179	0.6675	0.4382	0.6750	0.1690
$G_{\max}$	0.6584	0.2564	0.7223	0.1332	0.5439	0.3225	**0.6978*****	0.4013	0.6695	0.1686
$\mathcal{C}_{\text{chain}}$	**0.6603**	0.2573	0.7361	0.1331	0.5512	0.3175	0.6702	0.4165	0.6836	0.1621

Table 24: AUROC (↑) and ECE (↓) across the UNO-Bench splits with MiniCPM-generated answers. Avg is the unweighted mean of the four per-split values (pooled-union results in Appendix D). Best result per column in bold; asterisks as in Table 21, not attached to the derived Avg column.

Method	Avg	Audio	MC	MO	Visual
Majority Vote (Self-Consistency)	0.6865	0.8500	0.4135	0.7149	0.7677
Level-Based Conf.	0.6573	0.7918	0.4343	0.6652	0.7380
Self-Probing Conf.	0.6931	0.8311	**0.4815*****	0.6733	0.7866
Two-Step CoT Conf.	0.6695	0.8234	0.4692	0.6402	0.7453
UMPIRE	0.6968	**0.8816*****	0.4451	0.6621	**0.7985*****
S_{smooth}	0.6976	0.8630	0.4273	0.7310	0.7691
S_{dens}	0.6924	0.8597	0.4219	0.7191	0.7690
S_{goal}	0.6866	0.8463	0.4027	0.7265	0.7709
G_{avg}	**0.7015**	0.8608	0.4314	**0.7339*****	0.7798
$G_{\max}$	0.6913	0.8445	0.4265	0.7223	0.7719
$\mathcal{C}_{\text{chain}}$	0.6928	0.8548	0.4180	0.7232	0.7753

Table 25: AURAC (↑) across the UNO-Bench splits with MiniCPM-generated answers. Avg as in Table 24. Best result per column in bold; asterisks as in Table 21, not attached to the derived Avg column.

L GEOMETRY-WEIGHTED RESULTS WITH MINICPM

To complement the geometry-weighted results of Appendix K, we repeat that evaluation protocol using $n = 6$ candidate answers generated by MiniCPM instead of Gemini. As in the main text, MiniCPM being open-source means that UMPIRE here runs in its *default* configuration, computing internal states from the very model that generated the answers, rather than through the external proxy model required for closed models such as Gemini. All other settings are unchanged (means across 100 randomized evaluation runs; asterisks follow the 100-run t-test convention of Table 21). Tables 24–26 mirror Tables 21–23.

The MiniCPM results confirm the findings of Appendix K while sharpening the comparison with UMPIRE. First, the benefit of our geometric re-weighting over plain self-consistency is more uniform than with Gemini: the paired analysis (Table 26) shows significant relative gains on all four splits for AURAC (average $+1.24\%$, against -0.10% with Gemini) and on the Audio, multiple-choice, and Visual splits for AUROC (average $+0.68\%$); the single exception is AUROC on the open-answer omni-modal split ($-1.16\%^{**}$), where AURAC nevertheless improves ($+1.12\%^{***}$). Second, as expected, UMPIRE benefits substantially from its native configuration: with internal states from the generating model itself it becomes the strongest method on the unimodal Audio and Visual splits and achieves the best average calibration (ECE 0.1577), while G_{avg} attains the best average AURAC (0.7015, against 0.6968 for UMPIRE) and our aggregated score retains the best average AUROC in the table (0.6603). Crucially, however, its weakness in genuinely omni-modal settings persists: the open-answer omni-modal split remains UMPIRE's worst setting (AUROC 0.5875 vs. 0.6702 for our method), and paired tests confirm that $\mathcal{C}_{\text{chain}}$ outperforms it there by $+13.8\%$ AUROC*** and $+9.3\%$ AURAC***. This demonstrates that UMPIRE's difficulty in handling more than one extra-textual modality at a time is structural rather than an artifact of the proxy-model setting used for closed models—and that our method retains its omni-modal advantage

Split	ΔAUROC (%)	ΔECE (%)	ΔAURAC (%)
Audio	$+1.72^{***}$	$+1.61^{**}$	$+1.25^{***}$
MC	$+0.93^{***}$	-0.70^{***}	$+2.09^{***}$
MO	-1.16^{**}	$+2.87^{***}$	$+1.12^{***}$
Visual	$+1.21^{***}$	-1.93^{***}	$+0.51^{**}$
Average	+0.68	+0.46	+1.24

Table 26: Mean paired relative difference (%) of $\mathcal{C}_{\text{chain}}$ with respect to Self-Consistency across 100 runs with MiniCPM-generated answers. Positive ΔAUROC/ΔAURAC and negative ΔECE favor our method. Asterisks as in Table 23. The Average row is the unweighted mean of the four per-split deltas and carries no significance test; pooled-union deltas are reported in Appendix D.

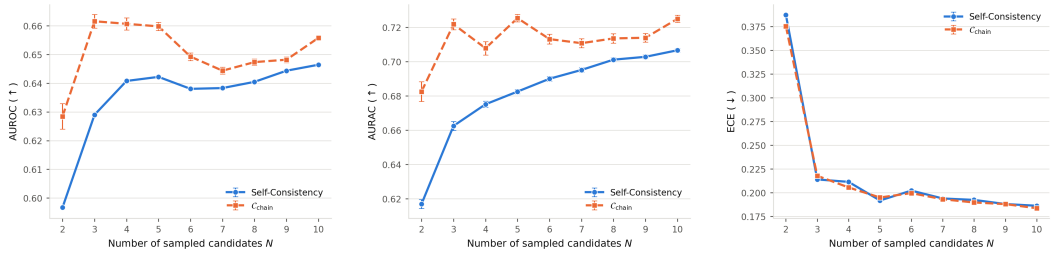


Figure 6: AUROC (↑), AURAC (↑), and ECE (↓) pooled over the union of splits (Appendix D) as a function of the number of sampled candidates N (MiniCPM-generated answers; means over 100 randomized evaluation runs per budget, error bars show ± 1 standard error of the mean).

while requiring no access to internal states whatsoever, and therefore applying unchanged to closed-source models where UMPIRE's default setting is simply unavailable. Finally, $G_{\max}$ behaves very differently across the two generators: it is the weakest signal on Gemini-Audio and the strongest single signal on MiniCPM-MO (0.6978, above Self-Consistency's 0.6782), which indicates that the discriminative power of individual embedding-based components depends on the answer-generating model and motivates selecting among them rather than fixing one a priori.

M Effect of the Number of Sampled Candidates

The sampling-based methods of this work fix the sampling budget at $N = 6$ candidates per question. To assess how the comparison between plain self-consistency and the geometry-weighted variant of Appendix K evolves with the sampling budget, we repeat that evaluation with MiniCPM-generated answers for $N \in \{2, \ldots, 10\}$, running 100 randomized evaluation repetitions per budget with the same protocol as the main experiments. Figure 6 reports AUROC, AURAC, and ECE pooled over the union of splits as a function of N for self-consistency and $\mathcal{C}_{\text{chain}}$. Pooled point estimates carry the heterogeneity caveat of Appendix D; we rely on them here only for the *paired* comparison between two methods evaluated on identical pools at each budget, for which the between-split component is shared.

Three observations emerge. First, $\mathcal{C}_{\text{chain}}$ exceeds self-consistency on AUROC and AURAC at every tested budget, while the two methods are essentially indistinguishable on ECE (differences within 0.012, of either sign). Second, and most practically relevant, the advantage of geometry-weighted aggregation is largest in the low-budget regime: at $N = 2$ the AUROC gap is $+3.2$ points (0.628 vs. 0.597) and the AURAC gap $+6.6$ points (0.683 vs. 0.617), and at $N = 3$ they are $+3.3$ and $+5.9$ points respectively. This is intuitive: with few samples, raw vote counts are extremely coarse

	Avg		Audio		MC		MO		Visual	
Method	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE
Self-Consistency	0.5927	0.3068	0.4512	0.6203	0.5630	0.2723	0.6334	0.2459	0.7234	**0.0888*****
Semantic Entropy	0.5882	0.3882	0.4471	0.6891	0.5531	0.3593	0.6319	0.3316	0.7207	0.1727
Num. Semantic Sets	0.5810	0.3106	0.4441	0.6237	0.5438	0.2914	0.6292	0.2274	0.7070	0.0998
RDS	0.6038	0.2405	0.6654	**0.2760*****	0.5434	**0.1382*****	0.5894	0.1951	0.6169	0.3526
Semantic Volume	0.6017	**0.2211**	**0.6671*****	0.4689	0.5417	0.1802	0.5918	**0.1223*****	0.6062	0.1130
$\mathcal{C}_{\text{chain}}$	0.6012	0.3014	0.4626	0.6107	**0.5631**	0.2744	**0.6517****	0.2269	0.7275	0.0937
$\mathcal{C}_{\text{chain}}$ + dispersion	**0.6431**	0.3474	0.6386	0.7205	0.5591	0.2964	0.6441	0.2474	**0.7304*****	0.1253

Table 27: AUROC (↑) and ECE (↓) of the additional black-box baselines across the UNO-Bench splits (Gemini-generated answers; means over 100 randomized evaluation runs). All methods in this table require the $N = 6$ sampled candidates. Avg is the unweighted mean of the four per-split values (pooled-union results in Appendix D). Bold indicates the best result per column; asterisks follow the 100-run paired t-test convention of Table 21 and are not attached to the derived Avg column.

(with $N = 2$, majority voting cannot distinguish most configurations at all), and the continuous embedding-based confidence provides exactly the tie-breaking signal that vote frequencies lack. Third, as N grows the two curves converge, with $\mathcal{C}_{\text{chain}}$ retaining a small but consistent edge (e.g., AUROC 0.656 vs. 0.646 at $N = 10$): vote frequencies become increasingly informative on their own, progressively subsuming the benefit of re-weighting. The variability of $\mathcal{C}_{\text{chain}}$ across runs is also larger at small N, reflecting the higher variance of fitting the aggregation weights on fewer effective votes. Taken together, these results indicate that our re-weighting is most valuable precisely where sampling is most expensive to scale—small candidate pools—which is the operative regime for costly omni-modal API models.

N Additional Baselines

To better isolate the contribution of our embedding-based signals from that of generic semantic-dispersion information between samples, we additionally compare against four established black-box baselines that, like self-consistency, operate on the N sampled answers; the two dispersion-based ones (RDS and Semantic Volume) are also reported in the main Tables 1 and 2, and this appendix provides their in-depth analysis in the Gemini setting. The first two use the semantic clusters of the sampled answers. **Semantic Entropy** (SE) follows the discrete formulation of Farquhar et al. (2024): given the cluster assignments produced by the same clustering step used throughout our experiments, the confidence is $\exp(-H)$, where H is the entropy of the empirical cluster distribution. **Number of Semantic Sets** (NumSets) uses the count of distinct clusters k as an uncertainty signal, mapped to a confidence as $1 - (k-1)/(N-1)$. The other two operate on continuous per-candidate *answer embeddings*, computed with the same E5-Omni encoder used for our scores: **RDS** (radial dispersion score) is the mean Euclidean distance of the N candidate answer embeddings from their centroid, and **Semantic Volume** is the log-determinant of the Gram matrix of the mean-centred embeddings, $\log\det(XX^\top + \varepsilon I)$ with $\varepsilon = 10^{-6}$; both are uncertainty scores, negated to obtain confidences. All four scores are question-level and are evaluated with the identical protocol, preprocessing, and 100-run repetition scheme of the geometry-weighted experiments of Appendix K (Gemini-generated answers); the Self-Consistency and $\mathcal{C}_{\text{chain}}$ rows are re-computed within the same runs and match the values of Appendix K up to run-to-run noise. We further evaluate a *hybrid* aggregate ($\mathcal{C}_{\text{chain}}$ + dispersion): the logistic-regression aggregator of $\mathcal{C}_{\text{chain}}$ refitted, under the same protocol, on seven features, i.e. the five chain-level scores plus the two dispersion scores added as question-level features. Table 27 reports AUROC and ECE.

Three findings emerge, one of which qualifies our earlier reading of this comparison. First, the discrete cluster-based baselines add nothing over plain self-consistency: SE tracks it closely on AUROC (0.5882 vs. 0.5927 on average) while degrading calibration, NumSets is uniformly weaker, and the same holds for AURAC (pooled values 0.8041 and 0.8002 vs. 0.8059; per-split AURAC was not stored for these baselines). For this family, self-consistency is effectively the best choice in its simplicity: with a modest sampling budget the cluster distribution carries essentially the same information as the majority-vote share, and re-parameterisations of it add nothing.

Second, and in contrast, the *continuous* embedding-dispersion baselines are considerably stronger than we anticipated. Semantic Volume attains the best average ECE in the table (0.2211), and both

RDS and Semantic Volume repair the Audio failure mode that affects every other sampling-derived method (AUROC 0.6654 and 0.6671^{***}, Audio AURAC up to 0.9017, where self-consistency sits below chance): the continuous dispersion of answer embeddings evidently carries signal on Audio that the discrete vote frequency inverts. Under the pooled-union protocol this Audio repair, combined with the between-split ranking effects analysed in Appendix D, even hands Semantic Volume the best pooled AUROC of the table; on the split average its discrimination is instead on par with $\mathcal{C}_{\text{chain}}$ (0.6017–0.6038 for the two dispersion baselines vs. 0.6012). Inter-sample dispersion measured in the embedding space is therefore a genuinely informative signal beyond cluster agreement, and the improvements of our method over self-consistency cannot be dismissed as generic dispersion information only where these baselines are also strong. Crucially, however, the dispersion baselines collapse exactly where our within-chain scores are strongest: $\mathcal{C}_{\text{chain}}$ retains a clear advantage on the Visual (AUROC 0.7275^{***} vs. 0.6169), multiple-choice (0.5631 vs. 0.5434), and open-answer omni-modal (0.6517^{***} vs. 0.5918) splits, with the corresponding AURAC gaps equally wide (e.g. MO 0.5193 vs. 0.4734). The two signal families are thus complementary rather than redundant: between-sample dispersion excels where answer diversity is informative (Audio), while the geometry of the reasoning chain excels where the difficulty lies in assessing an individual answer against multimodal evidence. We also note that RDS and Semantic Volume, like self-consistency and our weighted scheme, require the full set of N sampled candidates plus one embedding pass per candidate; unlike our method, however, they use only the answers' embeddings and discard the reasoning trace entirely. The MiniCPM-side results (Table 2) qualify the strength of this family further: with a generator whose agreement signal is not inverted on Audio, the dispersion baselines drop to the bottom of the sampling-based family on the split average (0.5764–0.5882) and fall to chance on both omni-modal splits, indicating that much of their Gemini advantage stems from repairing the Audio inversion.

Third, the hybrid aggregate gives an initial picture of how far simple feature-level fusion can combine the two families. Its dominant effect is to remove the worst-case failure of the chain-based aggregate: on Audio, where $\mathcal{C}_{\text{chain}}$ sits below chance, the hybrid reaches 0.6386 AUROC and 0.8529 AURAC (paired gains of $+0.176$ and $+0.038$ over $\mathcal{C}_{\text{chain}}$, both $p < 10^{-70}$), and it additionally attains the best Visual AUROC in the table (0.7304^{***}) and, as a direct consequence of the Audio repair, the best average AUROC (0.6431). No split is left at or below chance, which is true of no other method in this work. These gains come at a cost: small but significant losses against $\mathcal{C}_{\text{chain}}$ on pooled AURAC (-0.013), open-answer AUROC (-0.008), and calibration on every split, together with visibly larger run-to-run variance, which we attribute to the aggregator having to arbitrate between heterogeneous signal families from roughly 150 validation questions per split. In turn, the hybrid does not reach the per-split maxima of the specialised methods (Semantic Volume retains the best Audio AUROC and average ECE, $\mathcal{C}_{\text{chain}}$ the best open-answer AUROC). We read this as evidence that the complementarity is real and exploitable, and that extracting it fully requires a more careful fusion than feature concatenation (e.g. per-split gating between the two families), which we leave to future work.

O Reproducibility Details

O.1 Code and Data release

The code required to replicate the results in this work can be accessed in an anonymised form at `https://anonymous.4open.science/r/Multimodal-CoT-Confidence-Scoring-C12E`. The dataset used is, as mentioned, the publicly available UNO-Bench (Chen et al., 2025a), with further description of pre-processing and details for reproducing the data on which we run our experiments available in Appendix A.

O.2 Software Environment

All experiments were implemented in Python and executed using widely adopted open-source machine learning and scientific computing libraries. Table 28 summarizes the primary software dependencies used in our implementation.

All pretrained models and datasets were obtained through their official releases via the Hugging Face ecosystem or the corresponding model providers. Experiments were conducted using deterministic

Package	Version
Python	3.12
numpy	2.3.2
pandas	2.3.1
scipy	1.16.1
scikit-learn	1.7.1
torch	2.8.0
torchvision	0.23.0
transformers	4.56.0
huggingface-hub	0.34.4
sentence-transformers	5.1.0
accelerate	1.10.0
tokenizers	0.22.0
safetensors	0.6.2
google-genai	1.30.0
tqdm	4.67.1
Pillow	11.3.0
opencv-python	4.12.0
matplotlib	3.10.5
seaborn	0.13.2
PyYAML	6.0.2

Table 28: Primary software dependencies used in our implementation.

random seeds whenever applicable. The complete source code, configuration files, and dependency specifications will be released to facilitate reproducibility.