Matchability-Aware Conformal Prediction for Open-Ended Language Model Generation

Ilia Azizi

University of Lausanne, Switzerland | BegooAI, Switzerland | ilia@begoo.ai

NEURAL INFORMATION PROCESSING SYSTEMS

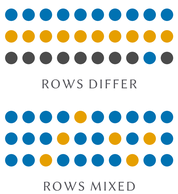
TL;DR

- CAELUM (Calibrated Aleatoric-Epistemic Language Uncertainty Modeling) builds semantic prediction sets for LLM answers while separating aleatoric from epistemic uncertainty.
- Aleatoric uncertainty targets input ambiguity. Epistemic uncertainty targets knowledge gaps, which can result in hallucinations.
- Coverage is capped by *matchability*, the fraction of items whose reference answer appears in any sampled cluster.
- Matchability can be predicted from the original question features using a lightweight classifier. We use those predictions to define a stratum and derive a conformal coverage bound within it.

1 Motivation

- Large language models (LLMs) often do not produce calibrated uncertainty estimates for their answers.
- Semantic entropy[1] and kernel language entropy[2] use uncertainty estimates to detect hallucinations, without decomposing uncertainty into aleatoric and epistemic components or constructing prediction sets with conformal coverage guarantees.
- Conformal prediction (CP) methods offer finite-sample marginal coverage guarantees for regression and classification tasks under exchangeability.
- Applying conformal methods to open-ended text generation is challenging due to the response space being large and the discrepancies between semantic and exact-match response accuracy.

2 Two sources of uncertainty



Aleatoric uncertainty U_{ale} is variation that remains given the inputs. For an LLM it arises from input ambiguity. For example, "What is the capital of South Africa?" has three valid answers depending on which branch of government the user means.

Epistemic uncertainty U_{epi} is limited knowledge of the predictive relation. For an LLM it reflects the model's lack of knowledge, for example where training data are sparse.

One clarification of x per row. Dots are sampled answers colored by semantic cluster. Disagreement between clarifications raises U_{ale}, and disagreement within a clarification raises U_{epi}.

3 Key terms

Semantic cluster	Sampled answers grouped by semantic equivalence. p_k is a cluster's share of the answers.
Prediction set $\mathcal{C}(x)$	A set of semantic clusters of sampled answers for query x.
Coverage	The reference answer's cluster is in $\mathcal{C}(x)$. The target is $\mathbb{P}(y^* \in \mathcal{C}(x)) \geq 1 - \alpha = 0.9$.
Set size (APSS)	Mean number of clusters per set. Lower is better at equal coverage, as smaller sets are more informative.
Matchability M	$M = 1$ when the reference answer appears in some sampled cluster, and $\pi = \mathbb{P}(M = 1)$.
Calibration	The set threshold $\hat{\tau}$ is fitted on held-out labelled questions (split conformal prediction).

4 Method

For a query x, we generate n clarifications and m LLM responses for each clarification[3], and cluster the nm responses by meaning. The uncertainty scores[4] are the pooled $U_{\text{total}}(x)$, the mean per-clarification $U_{\text{epi}}(x)$, and the residual $U_{\text{ale}}(x) = \max(U_{\text{total}} - U_{\text{epi}}, 0)$.

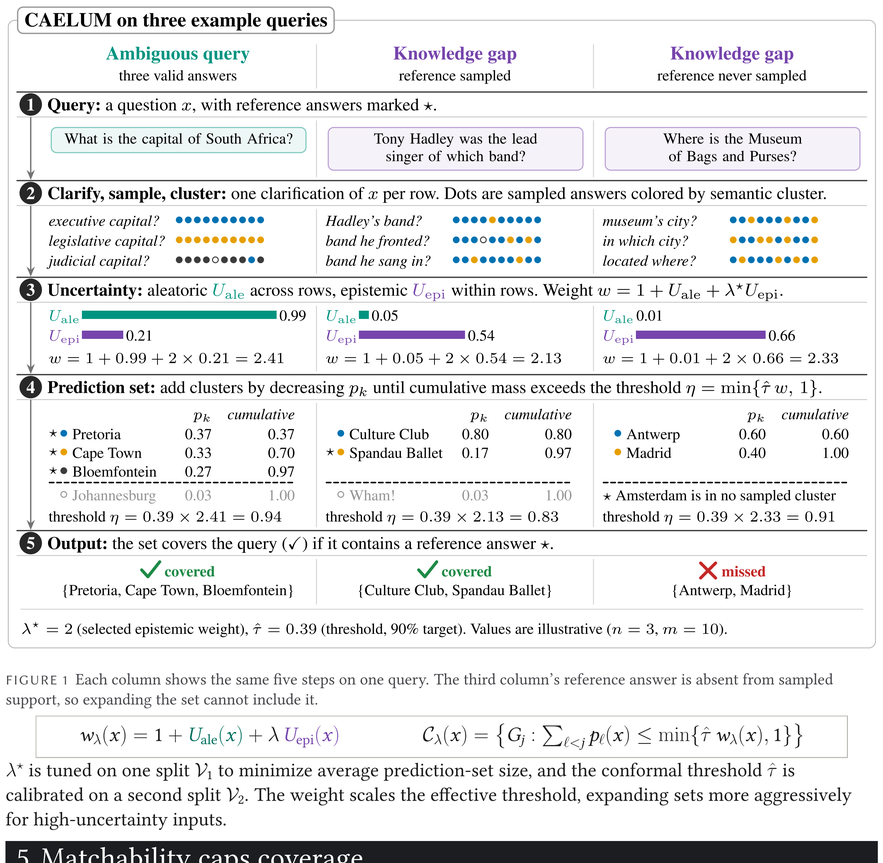
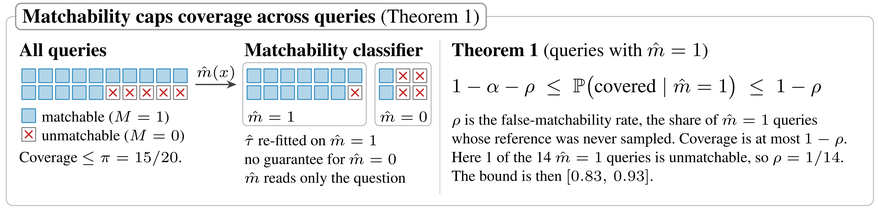


FIGURE 1 Each column shows the same five steps on one query. The third column's reference answer is absent from sampled support, so expanding the set cannot include it.

$$w_\lambda(x) = 1 + U_{\text{ale}}(x) + \lambda\, U_{\text{epi}}(x) \qquad \mathcal{C}_\lambda(x) = \left\{ G_j : \textstyle\sum_{\ell<j} p_\ell(x) \leq \min\{\hat{\tau}\, w_\lambda(x), 1\} \right\}$$

λ^* is tuned on one split $\mathcal{V}_1$ to minimize average prediction-set size, and the conformal threshold $\hat{\tau}$ is calibrated on a second split $\mathcal{V}_2$. The weight scales the effective threshold, expanding sets more aggressively for high-uncertainty inputs.

5 Matchability caps coverage



If $M = 0$, no threshold over the sampled clusters can include the reference, so population coverage is bounded above by $\pi = \mathbb{P}(M = 1)$ regardless of α. A classifier $\hat{m}(x)$ trained on question features before generation selects a predicted-matchable stratum. Calibrating only within it (Mondrian CP[5]) gives the coverage bound of Theorem 1. If $\rho > \alpha$, the target is unattainable within that stratum.

6 Support limits coverage

Six QA benchmarks (2,104 questions), three open-weight 4–9B LLMs, 50 answers per question.

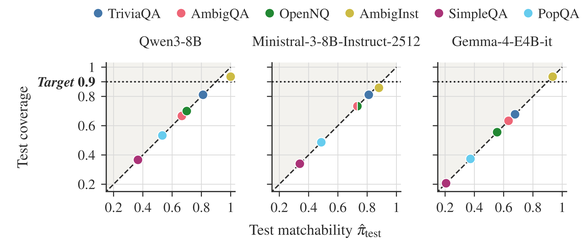


FIGURE 2 Test coverage and matchability for 18 model–dataset cells. The shaded region above the diagonal is unreachable using sampled clusters.

Test coverage for CP methods is close to the test support fraction in most cells. Once a set contains all sampled clusters, weighting cannot add a reference absent from those clusters.

7 Matchability can be predicted

Mean coverage	Learned stratum	Overall
30 combinations, 78 evaluations	0.74 ± 0.15	0.65 ± 0.17
18 cells, 4 splits	0.71 ± 0.14	0.61 ± 0.16

TABLE 1 Learned classifier only, without oracle fallback (78 and 45 evaluations). Generators range from 4B to 32B.

A classifier fit on one LLM's match labels predicts another LLM's labels on the same questions, with mean AUC 0.80 in-sample and 0.78 across generators. Observed label agreement is 0.81, against 0.50–0.53 for independent labels.

8 Flagging missing answers

U_{epi} ranks support failures better than U_{total} in 15 of 17 cells across four splits. A learned CAELUM detector adds an abstention label to the set less often than two baseline rules.

Trigger	AUROC ↑	Abstain ↓	Precision ↑	Aug. cov.
CAELUM detector	**0.780**	**0.459**	**0.569**	0.924
U_{epi} trigger	0.696	0.506	0.481	0.918
Good–Turing trigger[6]	0.550	0.878	0.371	0.989
CPQ module[6]	–	0.848	0.384	0.987

TABLE 2 29 cells with judged labels, four splits. Target augmented coverage 0.9.

9 Conclusion and limitations

- Across four random splits of 18 model–dataset combinations, matchability limits achievable coverage. Theorem 1 bounds coverage with a penalty term for false matchability.
- Larger generators do not remove the support limit. The pooled match rate rises from 61.1% (Qwen3-8B) to 66.0% (Qwen3-32B), with SimpleQA at 36.6%.
- CAELUM and the aleatoric-only baseline (Vanilla CP) have similar set sizes across splits, so the selected epistemic weight need not identify a particular uncertainty regime.
- Learned-stratum coverage on some factoid datasets, including PopQA, remains below the nominal target. Neither more samples nor escalation to larger models closes this gap.

REFERENCES

[1] Kuhn, Gal, Farquhar. Semantic uncertainty. ICLR 2023. [2] Nikitin et al. Kernel language entropy. NeurIPS 2024. [3] Hou et al. Input clarification ensembling. ICML 2024. [4] Walha et al. Fine-grained uncertainty decomposition in LLMs: a spectral approach. AAAI 2026. [5] Vovk et al. Mondrian confidence machine. 2003. [6] Noorani et al. Conformal prediction beyond the seen. NeurIPS 2025.