EPISTEMIC UNCERTAINTY IN LARGE LANGUAGE MODELS: MEASUREMENT, EVALUATION, MECHANISMS, AND APPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Token-level uncertainty in large language models (LLMs) arises from both knowledge gaps (epistemic uncertainty, EU) and the inherent flexibility of language (aleatoric uncertainty, AU). Accurate token-level EU estimation is fundamental to understanding model reliability at the basic decision unit of generation. However, the lack of a unified and direct testbed leaves estimator effectiveness, EU signal mechanisms, and downstream applicability insufficiently understood. To advance the understanding of uncertainty in LLMs, we make four contributions: **(1)** We introduce EU and AU proxy labels based on reasonable continuation sets and develop an automated pipeline to construct UncertainBench, comprising 7,216 labeled continuation positions across 15 domains. **(2)** We systematically evaluate EU across 67 LLMs and benchmark 35 uncertainty estimation methods, revealing that none combines accurate EU estimation with high EU purity. **(3)** We discover stronger and purer EU signals in the low-probability tail and propose a training-based hypothesis consistent with observed logit patterns. Emphasizing these signals yields a training-based estimator that outperforms baselines and generalizes across domains. **(4)** We apply EU estimators to five downstream tasks, helping distinguish reliable knowledge from lucky success, assess response quality, predict model capability, and identify critical branching positions. We also find that EU estimators offer limited value for predicting pre-trained model potential, whereas our proposed Rényi entropy-based method performs well by emphasizing signals of tokens in head probability buckets.

1 INTRODUCTION

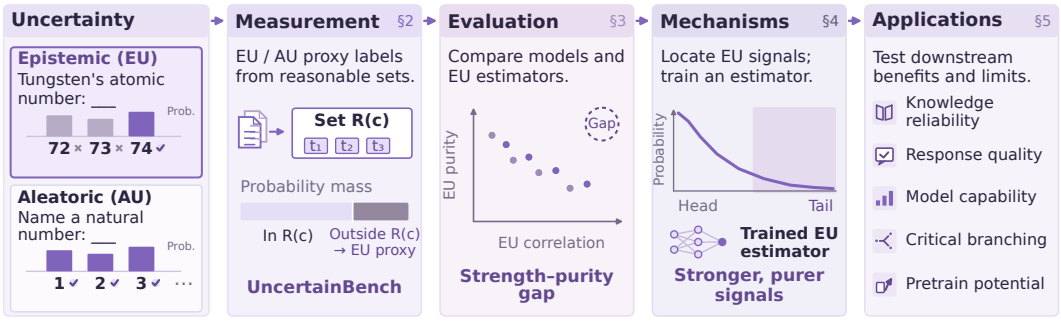


Figure 1: Overview of our study on token-level epistemic uncertainty.

The dispersion of an LLM's next-token distribution reflects its *total uncertainty* (TU), which can arise from both *aleatoric uncertainty* (AU) and *epistemic uncertainty* (EU) (Hüllermeier & Waegeman, 2021; Hou et al., 2024). AU is objective and captures the inherent multiplicity of reasonable continuations: the next-token distribution of "Name a natural number:" is naturally expected to have high entropy. EU is instead model-specific, stemming from gaps in the knowledge required by the context and potentially preventing the model from confidently predicting a particular next token.

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.

Accurate estimation of token-level EU[1] in LLMs has long been studied as a fundamental research problem (Baan et al., 2023; Geng et al., 2024), as it captures the reliability of a model at the basic decision unit of generation. Recent rapid advances in LLMs have increased the urgency of EU research, which may help understand and address hallucination, safety, and alignment challenges.

However, after systematically reviewing recent work on EU estimation, we are surprised to find that the field still lacks a direct and accurate testbed. Existing studies typically aggregate estimated token-level signals into response-level EU scores, then evaluate token-level estimators indirectly by correlating these scores with response correctness (Fadeeva et al., 2023; Geng et al., 2024). This indirect testbed has two serious limitations. First, the equivalence between more accurate response-level and token-level EU estimation remains unproven. Second, the absence of fine-grained token-level evaluation signals makes it difficult to identify estimator weaknesses and directions for improvement. The lack of a unified testbed further hinders systematic investigation into the mechanisms underlying EU signals in LLMs and the applicability limits of EU estimators.

In this work, we make four contributions to advance uncertainty estimation in LLMs (Figure 1):

We introduce EU and AU proxy labels and an automated pipeline for uncertainty benchmark construction. Prior work (Lahlou et al., 2023; Hou et al., 2024) adopts a reference-distribution view and defines AU as the entropy of an ideal oracle distribution $\bar{P}$ and EU as the discrepancy between the distribution P of the model under evaluation and $\bar{P}$. Since $\bar{P}$ is unobservable, we approximate it by a pooled distribution $\bar{P}_{\text{pool}}$ that assigns probability $1/|R(c)|$ to each token in a reasonable continuation set $R(c)$ for context c, and zero elsewhere. Substituting $\bar{P}_{\text{pool}}$ into the KL divergence and deriving a lower bound motivates the **EU proxy label** $\widehat{\text{EU}}(P, c) := \sum_{t \notin R(c)} P(t \mid c)$, *the probability that P generates an unreasonable next token*. The pooled oracle's entropy yields the **AU proxy label** $\widehat{\text{AU}}(c) := \log |R(c)|$. On this basis, we develop an automated pipeline for constructing uncertainty benchmarks with three steps: (1) selecting challenging continuation positions; (2) constructing candidate sets of reasonable tokens from three distinct sources; (3) establishing reasonable token sets through careful validation of each candidate token. An F1 score of 93.4% against manually constructed reasonable sets supports the pipeline's accuracy and coverage. Finally, we apply this pipeline to text collected from 2026 onward to construct **UncertainBench**, comprising **7,216 labeled continuation positions across 15 domains**.

We systematically evaluate EU across LLMs and benchmark existing uncertainty estimators on UncertainBench. We first examine 67 open-source LLMs through their unreasonable continuation rates, revealing a strong association with downstream performance and showing that post-training does not consistently reduce $\widehat{\text{EU}}$. We then evaluate 35 existing uncertainty estimation methods and denote the resulting Spearman correlations by ρ_{EU} and ρ_{AU} and define *EU purity* as $\pi_{EU} = |\rho_{EU}|/(|\rho_{EU}| + |\rho_{AU}|)$. Most estimators correlate with both labels, often more strongly with AU. Self-Certainty (Kang et al., 2025) (denoted as $\mathcal{M}_{\text{free}}$) achieves the highest baseline $|\rho_{EU}| = 0.401$, but its higher $|\rho_{AU}| = 0.417$ limits EU purity to $\pi_{EU} = 49.1\%$. Conversely, the Bayesian estimator achieves the highest purity ($\pi_{EU} = 75.9\%$) but a weaker EU correlation ($|\rho_{EU}| = 0.265$). **No evaluated baseline combines accurate EU estimation with high EU purity.**

We locate and explain uncertainty signals in LLMs and develop generalizable EU estimators. To identify stronger and purer EU signals, we group tokens by probability rank into exponentially sized buckets and correlate each bucket's geometric mean probability with $\widehat{\text{EU}}$ and $\widehat{\text{AU}}$. Beyond the highest-probability bucket, $|\rho_{EU}|$ generally rises toward the tail and plateaus, whereas $|\rho_{AU}|$ first rises and then falls, revealing **stronger and purer EU signals in the low-probability tail**. We hypothesize that knowledge reinforcement leaves a signature in tail suppression, consistent with the observed logit patterns. We further train a shallow regressor $\mathcal{M}_{\text{train}}$ on bucket-level distribution features, improving $|\rho_{EU}|$ by 15.2% over $\mathcal{M}_{\text{free}}$ and raising π_{EU} to 67.4%. The estimator also generalizes well to held-out domains.

We explore the applications and limits of EU estimators across five downstream tasks. (1) Distinguishing reliable knowledge from lucky success: when two models answer the same question correctly, $\mathcal{M}_{\text{train}}$ best identifies the one that answers correctly more consistently. **(2) Assessing**

[1]Unless otherwise specified, EU in this paper refers to token-level epistemic uncertainty.

response quality: $\mathcal{M}_{\text{free}}$ has a slight advantage in distinguishing correct from incorrect responses to the same question. **(3) Predicting model capability:** $\mathcal{M}_{\text{train}}$ achieves the strongest absolute Spearman correlation with benchmark accuracy. **(4) Identifying critical branching positions:** we propose an algorithm to identify positions where next-token choices strongly affect the correctness of the final answer. $\mathcal{M}_{\text{train}}$ predicts these positions more accurately, particularly those involving factual knowledge, suggesting potential for hallucination localization. **(5) Predicting pre-trained model potential:** in an exploratory evaluation, we find that head-probability signals predict post-training performance more strongly than the EU estimators. Based on this, we propose a Rényi entropy-based method that emphasizes these signals and attains the highest observed correlation.

2 Measurement: From Unobservable Oracles to UncertainBench

2.1 Uncertainty Definitions and Proxy Labels

Given context $c = (x_1, \ldots, x_{i-1})$, the model under evaluation predicts the next token at position i from a finite vocabulary $\mathcal{V}$ according to $P(\cdot \mid c)$. Let $\bar{P}(\cdot \mid c)$ denote the distribution of an ideal oracle with complete knowledge relevant to c. Following Hou et al. (2024) and Lahlou et al. (2023), TU is predictive entropy, AU is oracle entropy, and EU is excess predictive risk under log loss:

$$\mathrm{TU}(P, c) := H(P), \qquad \mathrm{AU}(c) := H(\bar{P}), \qquad \mathrm{EU}(P, c) := D_{\mathrm{KL}}(\bar{P} \| P). \tag{1}$$

These definitions recover the Bayesian decomposition: under a posterior over oracle distributions with mean P, the expected AU and EU sum to TU (Hüllermeier & Waegeman, 2021) (Appendix A.1). From an uncertainty perspective, EU directly measures the token-level gap between the evaluated model and the ideal oracle, making accurate EU estimation valuable for guiding LLM optimization and trustworthy use. However, $\bar{P}$ is unobservable in practice, a fundamental obstacle that has long hindered the construction of a unified testbed. To address this, we consider pooling reasonable next tokens into a nonempty set $R(c) \subseteq \mathcal{V}$ and approximating $\bar{P}$ uniformly on this set: $\bar{P}_{\text{pool}}(t \mid c) := \mathbf{1}_{\{t \in R(c)\}} / |R(c)|$. This pooling (1) avoids estimating unknown relative probabilities among candidate tokens, removing a potential source of systematic error, and (2) enables an EU proxy label with an intuitive probability interpretation via Jensen's inequality:

$$D_{\mathrm{KL}}(\bar{P}_{\text{pool}} \| P) = \frac{1}{|R(c)|} \sum_{t \in R(c)} \log \frac{1}{|R(c)| P(t \mid c)} \geq -\log \sum_{t \in R(c)} P(t \mid c). \tag{2}$$

Transforming this bound by $z \mapsto 1 - e^{-z}$ and taking the pooled entropy give the proxy labels

$$\widehat{\mathrm{EU}}(P, c) := \sum_{t \in \mathcal{V} \setminus R(c)} P(t \mid c), \qquad \widehat{\mathrm{AU}}(c) := H(\bar{P}_{\text{pool}}) = \log |R(c)|. \tag{3}$$

We define the EU proxy label as the probability of an unreasonable next token and the AU proxy label as annotated continuation diversity, $\log |R(c)|$.

2.2 Automated Benchmark Construction Pipeline for UncertainBench

Building on the definitions of $\widehat{\mathrm{EU}}$ and $\widehat{\mathrm{AU}}$, we design the corresponding automated benchmark construction pipeline around three questions (Figure 2; Appendix B.2 provides construction details).

(1) How to select challenging, discriminative continuation positions? Given seed texts, we use a heterogeneous pool of *source models* $\mathcal{M}_{\text{src}}$ to identify shared tokenizer boundaries. At each aligned position, we estimate EU for $\mathcal{M}_{\text{src}}$ using three representative methods. We shortlist positions with the highest mean EU across models and methods. For each shortlisted position with context c, we extract the top-K token set $T_K^{(m)}(c)$ from source model's next-token predictive distribution and merge these sets into $T_K(c) := \bigcup_{m \in \mathcal{M}_{\text{src}}} T_K^{(m)}(c)$. We further query *position judges* $\mathcal{M}_{\text{pos}}$ with c and $T_K(c)$ to select the most challenging positions with bounded reasonable-set sizes from this high-EU shortlist for tractable annotation, forming the position set $\mathcal{I}$.

(2) How to improve the coverage of $R(c)$? For each position in $\mathcal{I}$, we collect potentially reasonable tokens from three sources. We start with $T_K(c)$ and the continuation token in the original text, then query *expansion models* $\mathcal{M}_{\text{exp}}$ to propose plausible additions given c and initial candidates. We merge and deduplicate all candidates to form the candidate reasonable token pool $C(c)$.

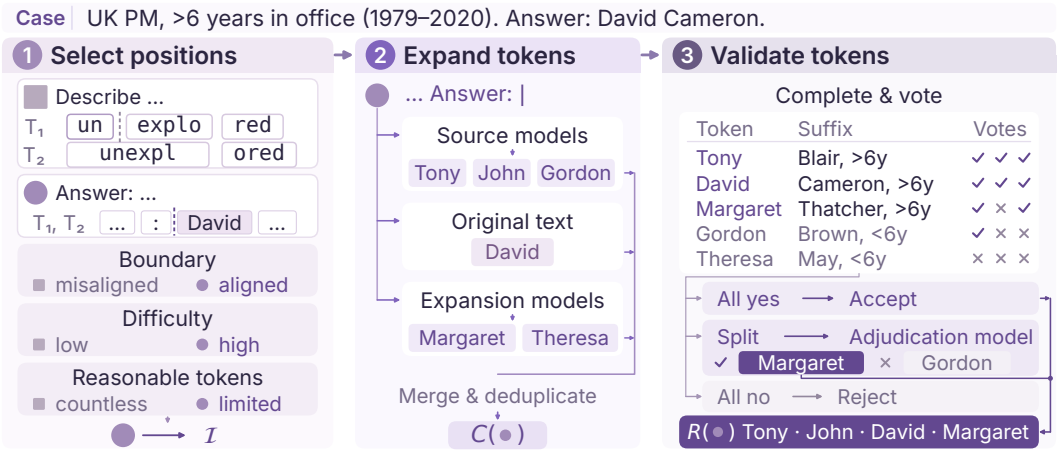


Figure 2: Automated Pipeline for Uncertainty Benchmark Construction.

(3) How to improve the precision of $R(c)$? For each token in $C(c)$, strong *validation models* $\mathcal{M}_{\text{val}}$ from different families assess its reasonableness by attempting to generate a suffix consistent with the context's linguistic and factual constraints. Unanimously approved tokens are added directly to $R(c)$. Tokens with split votes are then sent to a stronger *adjudication model* $\mathcal{M}_{\text{adj}}$, using the anonymized judgments and supporting evidence to make final decisions and establish $R(c)$.

Finally, we apply this pipeline to public texts collected from 2026 onward to construct **UNCERTAINBENCH**, comprising **7,216 labeled continuation positions from 3,608 texts across 15 domains**. The average reasonable-set size $|R(c)|$ is 49.91 tokens, with details in Appendix B. The established reasonable sets achieve an F1 score of 93.4% against manually constructed sets, based on **200 manual annotations**, supporting the reliability of both the pipeline and UNCERTAINBENCH. Each record contains a context, continuation position, and reasonable set. To evaluate a model, we obtain its next-token distribution at labeled positions and compute $\widehat{\text{EU}}(P, c)$ using equation 3.

3 EVALUATION: FROM LLM RELIABILITY TO ESTIMATOR EFFECTIVENESS

With reasonable sets established, UNCERTAINBENCH enables us to evaluate LLMs through EU proxy labels and examine how well existing uncertainty estimators recover these labels.

3.1 LLM CONTINUATION RELIABILITY

For each LLM, we average $\widehat{\text{EU}}$ over annotated positions to measure its unreasonable continuation rate (Figure 3). Our evaluation of 67 open-source LLMs yields five results and insights. Experimental details, full model rankings, and domain results are provided in Appendix C.1.

- **Leading models:** Kimi-K3 (Team, 2026), DeepSeek-V4-Pro-0813 (DeepSeek-AI, 2026), and Qwen3.8-2.4T-A95B (Qwen Team, 2026) achieve the lowest unreasonable rates of around 10%. These rates far exceed those observed at randomly selected positions during pipeline development ($< 1\%$), highlighting the effectiveness of challenging position selection.
- **Alignment with downstream performance:** mean $\widehat{\text{EU}}$ correlates slightly more strongly with downstream task performance (AA index v4.3.2[2]), achieving an absolute Spearman correlation of 0.698, versus 0.682 for perplexity (PPL), 0.678 for bits per byte (BPB), and 0.646 for entropy.
- **Knowledge-intensive domains are harder:** scientific texts generally elicit higher unreasonable continuation rates than code and news, with biomedical writing showing the highest mean rate.
- **Distinct label associations:** LLM-averaged $\widehat{\text{EU}}$ correlates weakly with $\widehat{\text{AU}}$ ($\rho = -0.121$) across positions, suggesting that UNCERTAINBENCH capture distinct signals related to EU and AU.
- **Mixed effects of post-training:** mean $\widehat{\text{EU}}$ increases after post-training in 9 of 23 base/post-trained model pairs. This suggests that reducing the unreasonable continuation rate relies primarily on language modeling and knowledge acquisition during pre-training.

[2]https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index

Model	Unreasonable continuations	Unreasonable continuation rate by domain $\widehat{\text{EU}}$ (%) ↓															PPL ↓	BPB ↓	AA ↑
		Scientific					Code					News	Wikipedia		Fiction				
	Avg. $\widehat{\text{EU}}$ (%) ↓	CS	Math	Other	Phys.	Bio.	C++	JS	MD	Other	Py.	BBC	En.	Other	En.	Other		bits/byte	v4.3
Kimi-K3	**9.78**	11.5	9.6	**11.1**	**11.4**	18.4	**5.1**	**4.9**	**8.4**	7.5	**6.5**	**5.3**	8.3	9.1	13.3	12.8	4.046	0.542	44
DeepSeek-V4-Pro-0813	**9.94**	**10.7**	9.9	11.2	11.8	**17.4**	6.5	5.9	9.0	8.1	9.1	5.9	**8.3**	7.2	**11.8**	11.3	6.633	0.733	36
Qwen3.8-2.4T-A95B	**10.46**	12.3	10.4	12.3	13.5	19.4	5.6	5.7	9.0	**7.2**	8.5	6.2	9.2	**7.1**	14.2	**10.1**	5.985	0.693	40
DeepSeek-V4-Pro-Base	**11.02**	12.6	10.7	12.1	13.0	19.2	6.6	6.5	11.4	9.4	9.3	6.6	9.7	7.9	13.0	11.9	4.778	0.606	—
DeepSeek-V4-Flash	**11.43**	13.0	11.6	12.9	13.1	19.0	7.7	7.2	13.1	9.4	9.5	5.8	9.8	8.7	11.9	10.6	9.298	0.864	35
Qwen3.5-397B-A17B	**11.51**	12.5	**9.5**	11.2	13.3	20.0	7.4	6.7	11.1	9.3	11.0	7.6	11.7	12.7	14.7	12.5	7.489	0.780	19
Kimi-K2.5	**11.91**	13.3	12.9	12.8	14.4	20.7	5.9	5.9	10.4	8.8	8.9	6.9	9.0	13.7	14.7	16.4	6.140	0.703	23
MiMo-V2.5-Pro-Base	**12.23**	13.2	14.0	14.5	14.3	21.3	6.1	6.6	11.7	8.3	8.9	6.5	10.4	12.1	14.1	15.8	5.539	0.663	—
GLM-5.3	**12.27**	14.1	14.1	13.9	14.1	21.2	6.9	6.4	10.6	9.6	9.8	6.1	10.2	10.4	14.7	15.3	8.714	0.839	45
DeepSeek-V3.2	**12.48**	14.7	12.3	13.4	13.4	20.7	7.4	7.2	12.9	10.5	10.6	7.2	10.1	13.2	14.1	15.7	7.180	0.764	16

Figure 3: Top-10 LLMs with the lowest unreasonable continuation rates on UNCERTAINBENCH. AA v4.3 denotes the Artificial Analysis Intelligence Index v4.3.2.

3.2 EFFECTIVENESS OF EXISTING UNCERTAINTY ESTIMATORS

Using the $\widehat{\text{EU}}$ and $\widehat{\text{AU}}$ proxy labels above, we further investigate what and how well 35 existing uncertainty estimators from four categories are actually measuring.

Method families. **Probability-based methods (P)** summarize the next-token distribution through its concentration, probability gaps, or token likelihoods (Geng et al., 2024; Kang et al., 2025; Zhang et al., 2024). **Methods based on model internals (I)** use internal activations or unnormalized logits to measure representation dispersion, prediction stability, or disagreement across layers and prediction heads (Liu et al., 2020; Chuang et al., 2024b; Chen et al., 2024). **Bayesian methods (B)** approximate uncertainty over predictive distributions through stochastic perturbations or evidence models, including methods that jointly estimate TU, AU, and EU (Gal & Ghahramani, 2016; Zhang et al., 2025; Ma et al., 2025). **Generative methods (G)** elicit confidence judgments through prompting (Kadavath et al., 2022). We name each estimator by its category and method abbreviation, such as P_{Shannon} for Shannon entropy; method details and complete results appear in Appendix C.2.

Evaluation criteria. For each model, we compute an estimator's Spearman correlations with $\widehat{\text{EU}}$ and $\widehat{\text{AU}}$, denoted by ρ_{EU} and ρ_{AU}. We also define *EU purity* as $\pi_{EU} := |\rho_{EU}|/(|\rho_{EU}| + |\rho_{AU}|)$, a measure of relative rank association rather than statistical independence.

Results and implications. Most estimators correlate with both labels, often more strongly with AU (Figure 4). In the available aggregate results, Self-Certainty P_{SelfCert} (denoted as $\mathcal{M}_{\text{free}}$) has the highest EU correlation ($|\rho_{EU}| = 0.401$), with AU correlation $|\rho_{AU}| = 0.417$ and purity 49.1%. Bayesian method $B_{\text{MC-EU}}$ achieves the highest EU purity ($\pi_{EU} = 75.9\%$), yet its EU correlation remains low ($|\rho_{EU}| = 0.265$). **These results indicate that no evaluated baseline combines strong EU estimation with high EU purity.** This gap motivates investigating the mechanisms underlying EU and AU signals and developing more effective EU estimators.

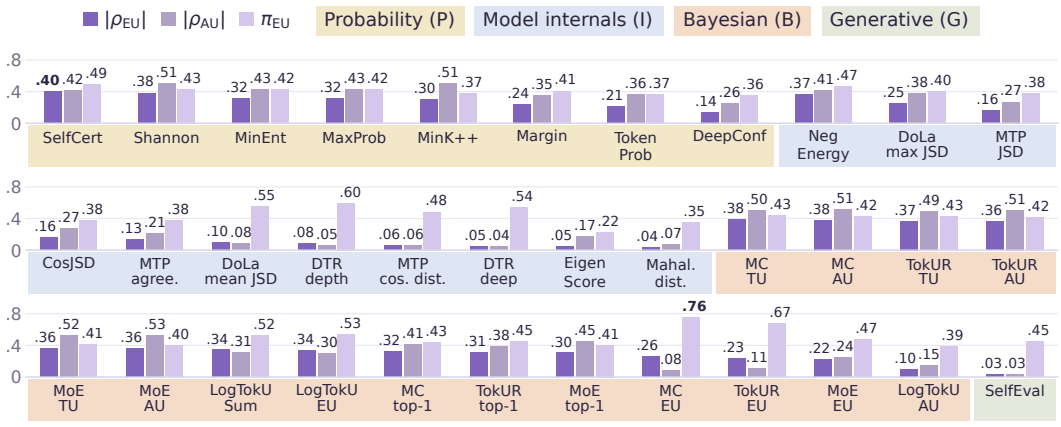


Figure 4: Absolute Spearman correlations of existing uncertainty estimators' predictions with $\widehat{\text{EU}}$ and $\widehat{\text{AU}}$ on UNCERTAINBENCH.

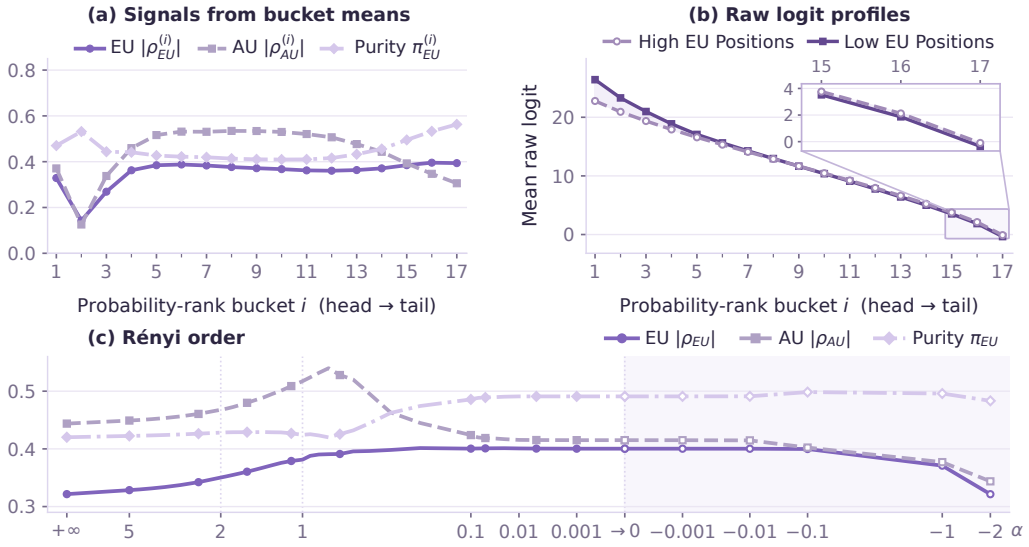


Figure 5: (a) EU/AU correlations of bucket mean log probabilities. (b) LLM-averaged mean raw logit profiles on high/low EU positions. (c) EU/AU correlations of Rényi entropy with varying α.

4 Mechanisms: From Signal Mining to Better Estimators

We seek uncertainty signals that generalize across LLMs by examining their probability distributions, a common representation across architectures. The leading performance of the probability-based method $\mathcal{M}_{\text{free}}$ suggests that probability distributions are a promising source of EU signals.

4.1 Locating and Explaining Uncertainty Signals

Signals across probability ranks. We sort next-token probabilities as $p_{(1)} \geq \cdots \geq p_{(|\mathcal{V}|)}$ and group the tokens into exponentially sized buckets. Each nonempty bucket B_i contains tokens ranked from 2^{i-1} to $\min(2^i - 1, |\mathcal{V}|)$. We use the mean log probability $s_i(P, c) := \frac{1}{|B_i|} \sum_{t \in B_i} \log P(t \mid c)$ as the bucket's probability signal. Correlating s_i with $\widehat{\text{EU}}$ and $\widehat{\text{AU}}$ across positions yields $|\rho^i_{EU}|$, $|\rho^i_{AU}|$, and EU purity π^i_{EU}. As shown in Figure 5a, beyond the first bucket, EU correlation rises and then dips before reaching a higher peak near the tail, whereas AU correlation declines after peaking in the middle. This yields the highest π^i_{EU} in the tail, suggesting that **tail probabilities help distinguish knowledge gaps from multiplicity of reasonable continuations**.

Mechanistic hypothesis and supporting evidence. To explain these findings, we hypothesize that repeated knowledge reinforcement leaves a cumulative signature in tail suppression. Individual updates need not strengthen every reasonable token, but repeated reinforcement can, on average, raise the logits of reasonable continuations while suppressing unreasonable tokens in the tail. We further hypothesize that this logit gap is proportional to the frequency with which the relevant knowledge is reinforced. Through softmax, a wider gap translates into stronger tail suppression relative to the head. **Lower tail probabilities may therefore indicate more frequent knowledge reinforcement and lower EU.** To examine this prediction, we divide positions into high- and low-EU groups separately for each LLM and compute each group's mean logit in every probability-rank bucket. The curves in Figure 5b cross: low-EU positions have higher head logits and slightly lower tail logits. The wider head-to-tail logit gap in these profiles supports the logit-level prediction of our hypothesis.

Explaining estimator performance through the observed mechanism. These findings also help explain baseline performance. Specifically, consider Rényi entropy $H_\alpha(P) := \frac{1}{1-\alpha} \log \sum_{t \in \mathcal{V}} P(t \mid c)^\alpha$ (Rényi, 1961). Self-Certainty is related to its first-order behavior near $\alpha = 0$, Shannon entropy is its $\alpha \to 1$ limit, and MaxProb is monotonically related to its $\alpha \to +\infty$ limit (Appendix D). For $H_\alpha(P)$, as α decreases from $+\infty$ through zero toward $-\infty$, effective weighting shifts from the

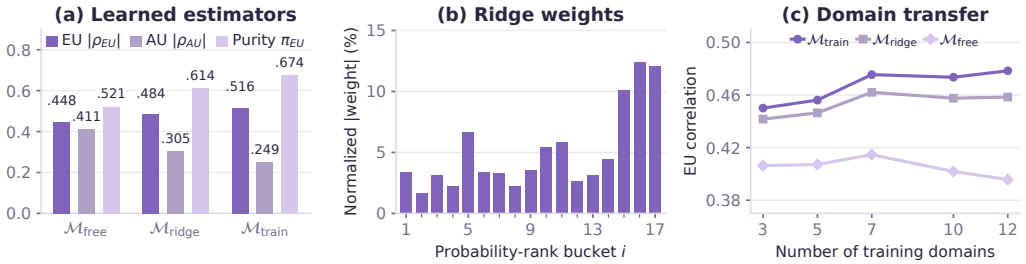


Figure 6: (a) EU/AU correlations and EU purity of $\mathcal{M}_{\text{free}}$, $\mathcal{M}_{\text{ridge}}$, and $\mathcal{M}_{\text{train}}$ on the same held-out test set. (b) Normalized absolute coefficients of $\mathcal{M}_{\text{ridge}}$ across probability buckets. (c) EU correlation of estimators on held-out domains as the number of training domains varies.

highest-probability token to all tokens equally, then to the lowest-probability token. As positive α approaches zero, the tail buckets contribute more to the signal as they contain most tokens, helping explain the overall rise in $|\rho_{EU}|$ (Figure 5c). As α decreases below zero, both EU and AU signals weaken, plausibly because overemphasizing a few extreme tail tokens amplifies noise.

4.2 Training-Based EU Estimator with Cross-Domain Generalizability

We next examine how much stronger and purer EU signals can be learned from distribution features.

Learning from distribution features. **(1) Ridge regression** $\mathcal{M}_{\text{ridge}}$ linearly combines standardized bucket mean log probabilities, yielding interpretable weights. **(2) A shallow nonlinear estimator with shared parameters**, $\mathcal{M}_{\text{train}}$, learns interactions among distribution features, including ranked log probabilities and bucket summaries (see Appendix D.4 for details). The network concatenates these features with an 8-dimensional model-specific embedding and then applies a 64-unit GELU hidden layer and a scalar linear output layer. This design, in which all LLMs share the backbone network, can provide more training signals and improve score comparability across LLMs.

Both $\mathcal{M}_{\text{ridge}}$ and $\mathcal{M}_{\text{train}}$ use $\widehat{\text{EU}}$ supervision alone, with UncertainBench split by source text into training, validation, and test sets in a $4:1:5$ ratio. On the same held-out test set, $\mathcal{M}_{\text{ridge}}$ improves on $\mathcal{M}_{\text{free}}$, and $\mathcal{M}_{\text{train}}$ further raises $|\rho_{EU}|$ to 0.516, a 15.2% improvement over $\mathcal{M}_{\text{free}}$ (Figure 6a). Without $\widehat{\text{AU}}$ supervision, $\mathcal{M}_{\text{train}}$ also reduces $|\rho_{AU}|$ from 0.411 to 0.249, raising π_{EU} from 52.1% to 67.4%. This suggests that $\mathcal{M}_{\text{train}}$ may improve $\widehat{\text{EU}}$ correlation by reducing interference from $\widehat{\text{AU}}$. Figure 6b shows that $\mathcal{M}_{\text{ridge}}$ assigns its largest normalized coefficient magnitudes to tail buckets, further supporting the view that these regions provide useful EU signals.

Transfer to other domains. To test whether an estimator learned on UncertainBench can be applied to other domains, we conduct a domain transfer experiment: we train on a subset of the 15 domains and evaluate on disjoint held-out domains without refitting. As shown in Figure 6c, $\mathcal{M}_{\text{train}}$ outperforms baselines across all tested training-domain counts. This supports a practical workflow: given a set of LLMs, train $\mathcal{M}_{\text{train}}$ using their distributions and $\widehat{\text{EU}}$ labels on UncertainBench; the resulting estimator is expected to be effective for these LLMs on new corpora.

5 Applications: From Upstream Signals to Downstream Tasks

Having identified the best-performing training-free and training-based estimators, we now go a step further by exploring their strengths and limitations across downstream tasks.

5.1 Assessing Knowledge Reliability, Response Quality, and Model Capability

We first average token-level EU estimates within each response and explore applications of the resulting response-level EU signals. We experiment with 32 LLMs across 31 benchmarks, sampling

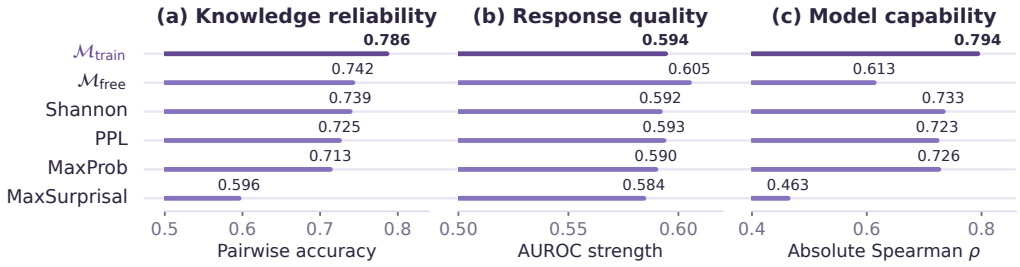


Figure 7: (a) Identifying the model with higher avg@k from correct responses. (b) Response correctness discrimination using AUROC strength (Appendix E.1). (c) Predicting benchmark accuracy from model-mean uncertainty.

$k = 10$ responses per question at temperature 1. Besides $\mathcal{M}_{\text{train}}$ and $\mathcal{M}_{\text{free}}$, we also include many estimators and list the best-performing ones in Figure 7. See Appendix E.1 for experimental details.

Distinguishing reliable knowledge from lucky success. When two models answer the same question correctly, we want to know which has a firmer grasp of the underlying knowledge, as indicated by a higher avg@k. As shown in Figure 7a, $\mathcal{M}_{\text{train}}$ selects the model with higher avg@k with 78.6% accuracy. This suggests that EU can help identify the more reliable model for a given query from one correct response per model, without sampling ten responses from each.

Assessing response quality. For the same model and question, we test whether EU distinguishes correct from incorrect responses. $\mathcal{M}_{\text{free}}$ slightly outperforms $\mathcal{M}_{\text{train}}$ in benchmark-mean AUROC strength (0.605 versus 0.594; Figure 7b). This reversal shows that stronger token-level EU estimation need not yield better response-level correctness discrimination, reinforcing the need for direct token-level evaluation and the construction of UNCERTAINBENCH.

Predicting model capability. Aggregating EU across responses yields a model-level score, which we correlate with accuracy within each benchmark. Averaged across benchmarks, $\mathcal{M}_{\text{train}}$ attains an absolute Spearman correlation of 0.794, exceeding Shannon entropy (0.733), MaxProb (0.726), PPL (0.723), and $\mathcal{M}_{\text{free}}$ (0.613; Figure 7c). For LLMs represented during estimator training, these results suggest an association between mean EU and model capability.

5.2 IDENTIFYING CRITICAL BRANCHING POSITIONS

Beyond assessing complete responses, we ask whether EU identifies token positions where next-token choices affect final-answer correctness. Let $Y(t, c)$ be the probability of a correct final answer after choosing token $t \in \mathcal{V}$ in context c. We define the *branching degree* as $B(c) := \sqrt{\sum_{t\in\mathcal{V}} P(t \mid c)(Y(t,c) - \bar{Y}(c))^2}$, where $\bar{Y}(c) = \sum_{t\in\mathcal{V}} P(t \mid c)Y(t, c)$. A large $B(c)$ indicates a critical branching position: the next-token choice strongly affects eventual correctness.

We use SimpleQA Verified (Haas et al., 2025) for its knowledge-intensive factual questions, which facilitate detailed analysis and categorization of branching decisions along reasoning trajectories.

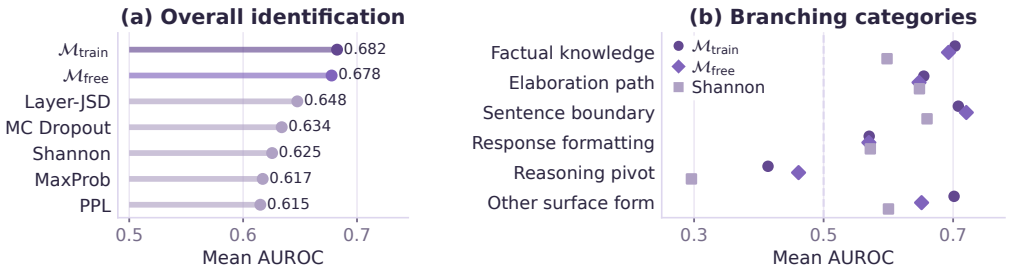


Figure 8: Mean AUROC for detecting labeled branching positions: (a) overall; (b) by category.

To locate critical branching positions efficiently, we design a progressive binary search that refines candidate intervals using Monte Carlo estimates of changes in continuation correctness, a path-based proxy for $B(c)$ (see Appendix E.2 for details). We evaluate 189 responses from Qwen3.6-35B-A3B with 913 labeled positions, at a search cost of approximately 6,200 suffix rollouts per labeled position. Treating detected path-change positions as positives and unmarked positions as negatives, we compute AUROC within each response. As shown in Figure 8a, $\mathcal{M}_{\text{train}}$ and $\mathcal{M}_{\text{free}}$ attain mean AUROCs of 0.682 and 0.678, respectively, outperforming the other baselines. We further categorize critical branching positions by their roles in reasoning and evaluate estimator performance in each category. On factual-knowledge positions, $\mathcal{M}_{\text{train}}$ reaches 0.703, compared with 0.598 for Shannon entropy (Figure 8b), suggesting potential for localizing knowledge-related hallucinations. All estimators perform relatively poorly at response-formatting and reasoning-pivot positions. We hypothesize that this is because EU estimators mainly reflect confidence in the current step and may fail to anticipate how a token redirects subsequent reasoning and the final answer.

5.3 Predicting Pre-Trained Model Potential

Finally, we explore whether uncertainty signals can predict a pre-trained model's performance after post-training, a central question in pre-training evaluation. We choose nine continued-pretraining checkpoints from internal development with different data mixtures, hyperparameters, and training-data volumes. Potential labels are their SWE-bench Verified (Jimenez et al., 2024), Pro (Deng et al., 2025), and Multilingual (Yang et al., 2025) scores after identical post-training. Using 7,597 trajectories generated by Claude Opus 4.6 (Anthropic, 2026) on these benchmarks, we average each checkpoint's uncertainty estimates over model-generated tokens and correlate them with potential labels across checkpoints. The results are shown in Figure 9b.

On this task, $\mathcal{M}_{\text{train}}$ and $\mathcal{M}_{\text{free}}$ attain mean absolute Spearman correlations of 0.834 and 0.741 across benchmarks, below the other baselines. We therefore examine individual probability-rank buckets. As shown in Figure 9a, bucket 3 has the strongest signal: lower mean log probabilities of its head tokens predict greater potential ($\rho = -0.939$), whereas top-1 log probabilities correlate positively ($\rho = 0.906$). Based on these observations, we focus on the cleanest signals by computing Rényi entropy over the renormalized top-k probabilities, using $0 < \alpha < 1$ to downweight the top-1 token. With $\alpha = 0.5$ and $k = 128$, the averaged score reaches $\rho = -0.944$, the largest observed magnitude in this exploratory hyperparameter scan. See Appendix E.3 for the hyperparameter scan.

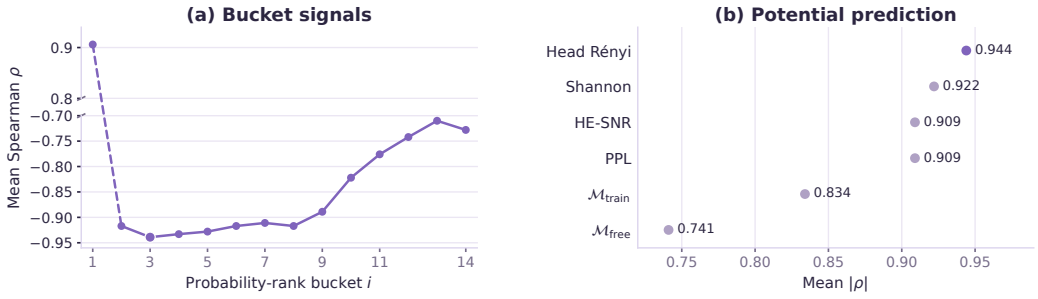


Figure 9: Potential prediction task: (a) signed bucket correlations (broken axis); (b) mean $|\rho|$ across three SWE benchmarks. Head Rényi uses $\alpha = 0.5$ and $k = 128$.

6 Conclusions

In this work, we study token-level epistemic uncertainty in LLMs, an area where progress has been hindered by the lack of a unified, direct benchmark. Our study brings together four complementary efforts: **(1) Measurement:** we introduce EU and AU proxy labels and construct UncertainBench; **(2) Evaluation:** we use this testbed to systematically assess 67 LLMs and 35 uncertainty estimators; **(3) Mechanisms:** we analyze uncertainty signals in predictive distributions and use these insights to develop a stronger EU estimator; **(4) Applications:** we examine the strengths and limitations of these estimators across five downstream tasks. We discuss the limitations of this study and opportunities for further research in Appendix F.

AI USE STATEMENT

Large language models assisted candidate expansion, validation, and adjudication in benchmark construction, as described in Section 2.2 and Appendix B.2. AI tools also assisted manuscript editing and preparation of visualization code. The authors are responsible for the final manuscript and research claims.

REFERENCES

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. Distinguishing the knowable from the unknowable with language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 503–549. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ahdritz24a.html.

Anthropic. Introducing Claude Opus 4.6, February 2026. URL https://www.anthropic.com/news/claude-opus-4-6.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. Uncertainty in natural language generation: From theory to applications. *CoRR*, abs/2307.15703, 2023. doi: 10.48550/ARXIV.2307.15703. URL https://doi.org/10.48550/arXiv.2307.15703.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 1419–1436. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.EMNLP-MAIN.84. URL https://doi.org/10.18653/v1/2024.emnlp-main.84.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=Th6NyL07na.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/ARXIV.2606.19348. URL https://doi.org/10.48550/arXiv.2606.19348.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean Hendryx, Zifan Wang, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. Swe-bench pro: Can AI agents solve long-horizon software engineering tasks? *CoRR*, abs/2509.16941, 2025. doi: 10.48550/ARXIV.2509.16941. URL https://doi.org/10.48550/arXiv.2509.16941.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 5050–5063. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.276. URL https://doi.org/10.18653/v1/2024.acl-long.276.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Lm-polygraph: Uncertainty estimation for language models. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pp. 446–461. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-DEMO.41. URL https://doi.org/10.18653/v1/2023.emnlp-demo.41.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024. doi: 10.1038/S41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1050–1059. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 6577–6595. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.366. URL https://doi.org/10.18653/v1/2024.naacl-long.366.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *CoRR*, abs/2509.07968, 2025. doi: 10.48550/ARXIV.2509.07968. URL https://doi.org/10.48550/arXiv.2509.07968.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19023–19042. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/hou24b.html.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506, 2021. doi: 10.1007/S10994-021-05946-3. URL https://doi.org/10.1007/s10994-021-05946-3.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know *When* language models know? on the calibration of language models for question answering. *Trans. Assoc. Comput. Linguistics*, 9:962–977, 2021. doi: 10.1162/TACL_A_00407. URL https://doi.org/10.1162/tacl_a_00407.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/ARXIV.2207.05221. URL https://doi.org/10.48550/arXiv.2207.05221.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/1c7eff166a8e345f664f0faa8f4e4d2e-Abstract-Conference.html.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5574–5584, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: direct epistemic uncertainty prediction. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=eGLdVRvvfQ.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with evidence. *arXiv preprint arXiv:2502.00290*, 2025.

Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.741. URL https://doi.org/10.18653/v1/2023.emnlp-main.741.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Kimi Team. Kimi K3: open frontier intelligence. *CoRR*, abs/2607.24653, 2026. doi: 10.48550/ARXIV.2607.24653. URL https://doi.org/10.48550/arXiv.2607.24653.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. 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, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.330. URL https://doi.org/10.18653/v1/2023.emnlp-main.330.

Yueyang Wang, Jiawei Fu, Baolong Bi, Xili Wang, and Xiaoqing Liu. HE-SNR: uncovering latent logic via entropy for guiding mid-training on SWE-BENCH. *CoRR*, abs/2601.20255, 2026. doi: 10.48550/ARXIV.2601.20255. URL https://doi.org/10.48550/arXiv.2601.20255.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 2734–2744. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EACL-MAIN.236. URL https://doi.org/10.18653/v1/2021.eacl-main.236.

John Yang, Kilian Lieret, Carlos E. Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. Swe-smith: Scaling data for software engineering agents, 2025. URL https://arxiv.org/abs/2504.21798.

Jingyang Zhang, Jingwei Sun, Eric C. Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Helen Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. volume abs/2404.02936, 2024. doi: 10.48550/ARXIV.2404.02936. URL https://doi.org/10.48550/arXiv.2404.02936.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris N. Metaxas, and Hao Wang. Token-level uncertainty estimation for large language model reasoning. *CoRR*, abs/2505.11737, 2025. doi: 10.48550/ARXIV.2505.11737. URL https://doi.org/10.48550/arXiv.2505.11737.

A Uncertainty Definitions and Proxy Label Derivations

This appendix establishes the relationship between the oracle-based definitions, the Bayesian decomposition, and the proxy labels used in Section 2.1.

A.1 Oracle and Bayesian uncertainty

For a fixed oracle $\bar{P}$, the expected log loss (using natural logarithms throughout) decomposes as (Lahlou et al., 2023)

$$\mathbb{E}_{t\sim\bar{P}}[-\log P(t \mid c)] = H(\bar{P}) + D_{\mathrm{KL}}(\bar{P}\|P), \tag{4}$$

so oracle EU is the model's excess risk above the irreducible oracle entropy. Given training data $\mathcal{D}$, let $q(\theta \mid \mathcal{D})$ be a posterior over model parameters Θ and $P_{\mathrm{B}}(\cdot \mid c) = \mathbb{E}_{\theta\sim q}[P_\theta(\cdot \mid c)]$ its posterior predictive distribution. The expected divergence from each conditional model to this posterior predictive satisfies

$$\begin{aligned}\mathbb{E}_{\theta\sim q}[D_{\mathrm{KL}}(P_\theta\|P_{\mathrm{B}})] &= \mathbb{E}_{\theta\sim q}\left[\sum_{t\in\mathcal{V}} P_\theta(t \mid c)\log\frac{P_\theta(t \mid c)}{P_{\mathrm{B}}(t \mid c)}\right] \\ &= H(P_{\mathrm{B}}) - \mathbb{E}_{\theta\sim q}[H(P_\theta)].\end{aligned} \tag{5}$$

This gives the Bayesian decomposition (Hüllermeier & Waegeman, 2021; Hou et al., 2024)

$$H(P_{\mathrm{B}}) = \underbrace{\mathbb{E}_{\theta\sim q}[H(P_\theta)]}_{\mathrm{AU_B}} + \underbrace{\mathbb{E}_{\theta\sim q}[D_{\mathrm{KL}}(P_\theta\|P_{\mathrm{B}})]}_{\mathrm{EU_B}}, \tag{6}$$

where $\mathrm{EU_B} = I(T;\Theta \mid c,\mathcal{D})$ for the next-token random variable T under the Bayesian joint distribution. If the unknown oracle is modeled as $\bar{P} = P_\Theta$ with posterior q, averaging AU and EU in equation 1 recovers $\mathrm{AU_B}$ and $\mathrm{EU_B}$ when $P = P_{\mathrm{B}}$. For a fixed oracle, systematic predictive mismatch may remain even when posterior models agree, as discussed by Lahlou et al. (2023) under model misspecification.

A.2 KL lower bound and proxy labels

For the uniform reference on a nonempty $R(c)$, Jensen's inequality gives

$$\begin{aligned}D_{\mathrm{KL}}(\bar{P}_{\mathrm{pool}}\|P) &= \frac{1}{|R(c)|}\sum_{t\in R(c)}\log\frac{1}{|R(c)|P(t \mid c)} \\ &= -\log|R(c)| - \frac{1}{|R(c)|}\sum_{t\in R(c)}\log P(t \mid c) \\ &\geq -\log|R(c)| - \log\left(\frac{1}{|R(c)|}\sum_{t\in R(c)}P(t \mid c)\right) \\ &= -\log\sum_{t\in R(c)}P(t \mid c).\end{aligned} \tag{7}$$

The increasing map $z \mapsto 1-e^{-z}$ transforms this lower bound into $1-\sum_{t\in R(c)}P(t \mid c) = \widehat{\mathrm{EU}}(P,c)$, while the pooled entropy is

$$H(\bar{P}_{\mathrm{pool}}) = -\sum_{t\in R(c)}\frac{1}{|R(c)|}\log\frac{1}{|R(c)|} = \log|R(c)|. \tag{8}$$

For $P(R(c) \mid c) > 0$, define $P_R(t \mid c) = P(t \mid c)/P(R(c) \mid c)$ on $R(c)$; then

$$D_{\mathrm{KL}}(\bar{P}_{\mathrm{pool}}\|P) = -\log P(R(c) \mid c) + D_{\mathrm{KL}}(\bar{P}_{\mathrm{pool}}\|P_R). \tag{9}$$

Thus the bound is tight when P_R is uniform, and the EU proxy label retains only the term determined by the probability assigned to $R(c)$.

A.3 Interpretation and scope of the labels

What the EU proxy label measures. Equation 9 separates probability assigned outside the reasonable set from the allocation of probability within it. The label $\widehat{\mathrm{EU}}$ measures the former and does not penalize preferences among reasonable alternatives. This makes the label invariant to reallocating a fixed probability mass among reasonable alternatives. It consequently ranks models by unreasonable continuation risk, which need not preserve their ordering under the full oracle KL.

What the AU proxy label measures. Among distributions supported on a set of size r, the uniform distribution has maximum entropy $\log r$. Consequently, $\widehat{\mathrm{AU}}$ measures the breadth of the annotated continuation space without estimating the oracle's relative preferences within that space. It equals the oracle entropy when the oracle is uniform on the annotated support; otherwise it is the entropy of the pooled reference. The label is zero for a singleton reasonable set.

Boundary cases and incomplete annotation. The pooled reference and its entropy require $R(c) \neq \varnothing$. If a reasonable token has zero model probability, the full KL is infinite under its usual extended definition. The support-mass lower bound remains finite when some other reasonable token has positive probability, and is infinite when the total reasonable mass is zero.

If annotation misses reasonable tokens, their probability is counted as unreasonable and the estimated support size decreases. If an unreasonable token is included, the effects reverse. This motivates both candidate expansion and validation: greater coverage alone does not guarantee more accurate labels. Across tokenizers, shared character boundaries align the conditioning text, but reasonable-set sizes still depend on token granularity and decoded-string matching.

B Benchmark Construction Details

We detail the corpus, annotation protocol, and quality checks underlying the reasonable continuation sets in Section 2.2.

B.1 Corpus, positions, and annotation units

The source collection contains 7,500 public texts, with 500 texts per domain. The retained benchmark contains 3,608 distinct texts and 7,216 continuation positions, spanning research articles, source code, repository documentation, news, encyclopedia entries, and fiction. Table 1 reports the domain breakdown. Text counts refer to distinct retained source texts; set sizes and their overall mean are computed over positions, without giving small domains extra weight.

Each annotated position contains a conditioning prefix, domain, original continuation token, candidate judgments, and final reasonable set. Shared tokenizer boundaries ensure that source models receive the same prefix, while explicit decoding checks reject misaligned boundaries. They do not require different tokenizers to produce identical next-token strings. The retained texts contain between 1 and 15 labeled positions: 1,804 texts contribute one, 894 contribute two, and the remaining 910 contribute three or more. Figure 10 summarizes this distribution and reasonable-set sizes.

B.2 Pipeline roles and annotation protocol

Position screening. The source pool provides tokenizer alignment, uncertainty scores, and seed predictions. We score candidate positions using Monte Carlo dropout EU (MC EU), perplexity (PPL), and negative energy (NegEnergy), standardize scores within each source text, and average them across source models and methods. Position judges assess four properties: coherent context, a sufficiently constrained continuation space, discriminative difficulty, and a need to understand the context. In particular, open-ended positions with an impractically large set of reasonable alternatives are rejected. This selection targets informative continuation decisions rather than a random sample of all generation positions. Based on the estimator comparison in Section 3.2, we recommend replacing PPL with Self-Certainty in future iterations of the construction pipeline.

Table 2 lists the source and annotation models by pipeline role.

Table 1: Domain statistics of UncertainBench. Text counts cover texts with at least one retained position; mean $|R(c)|$ is computed over positions.

Domain	Content	# Texts	# Positions	Mean $\lvert R(c)\rvert$
arXiv (CS)	Computer science papers	294	566	52.42
arXiv (Math)	Mathematics papers	367	747	36.76
arXiv (Physics)	Physics papers	378	787	48.15
arXiv (Other)	Other research papers	317	591	48.51
bioRxiv	Biology preprints	335	679	43.17
GitHub (C++)	C++ code	269	711	50.74
GitHub (JavaScript)	JavaScript code	235	486	55.79
GitHub (Markdown)	Repository documentation	194	404	55.34
GitHub (Other)	Other code	267	568	41.51
GitHub (Python)	Python code	237	406	48.94
BBC News	News articles	111	159	62.03
Wikipedia (English)	English encyclopedia	148	257	63.89
Wikipedia (non-English)	Non-English encyclopedia	178	390	51.24
AO3 (English)	English fiction	173	243	81.79
AO3 (non-English)	Non-English fiction	105	222	54.09
Overall	—	**3,608**	**7,216**	**49.91**

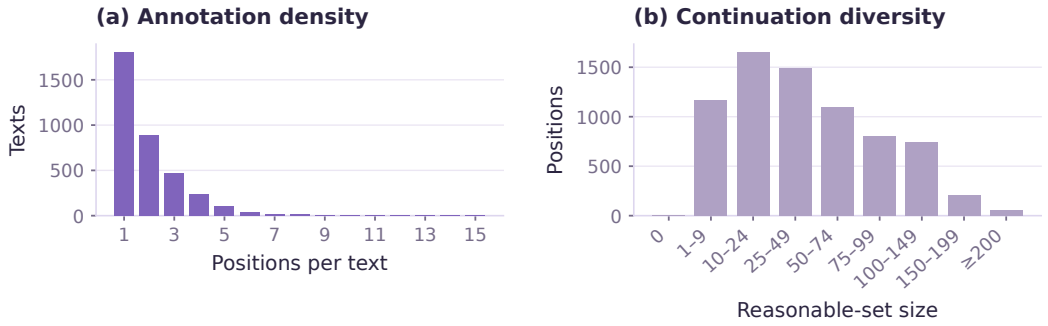


Figure 10: Annotation density and reasonable-set sizes in the final benchmark. Text counts are based on distinct source identifiers; set sizes use the stored annotations.

Candidate coverage. Seed extraction uses the top 60 predictions per source model, merges duplicate decoded strings, and combines them with the original continuation and expansion proposals. The protocol filters special and whitespace-only tokens and applies a seed-probability threshold of 0.001. For screening, each source model's retained probabilities are normalized before averaging across models. These weights guide selection; they are not the probabilities of the uniform reference used to define the labels. Expansion models receive the prefix and existing candidates, and propose short, nonduplicate continuations without copying the end of the prefix. The protocol flags candidate pools exceeding 300 entries as too open for annotation. Original-text tokens undergo the same validation as proposed tokens.

Validation and adjudication. Each validation model judges a candidate and provides supporting evidence: a short continuation for an accepted candidate or an explanation for rejection. Unanimous decisions are retained; disagreements are sent to the adjudication model. The adjudicator receives the context, candidate, decisions, and evidence under anonymous judge labels. Thus a provisional majority does not override a later adjudication, and the process retains the evidence behind disputed labels.

Decoded-string matching. Equations 2 and 3 describe the construction in a fixed token space. For cross-tokenizer scoring, annotations are stored as shared decoded strings $R_{\mathrm{str}}(c)$, inducing the

Role	Model identifiers
Source models	`dsv4_flash`, `hy3`, `glm47`, `kimi_k25`, `minimax_m3`, `llama4_scout`, `seed_oss_36b`, `step35_flash`, `qwen36_35b_a3b`
Position screening	`qwen3.7-max`
Candidate expansion	`gpt-5.4`, `glm-5.1`, `claude-opus-4-8`
Candidate validation	`claude-fable-5`, `kimi-k2.6`, `qwen3.7-max`
Dispute adjudication	`claude-opus-5`

Table 2: Source and annotation models used in benchmark construction.

accepted set $A_m(c) := \{t \in \mathcal{V}_m : \mathrm{match}(\mathrm{decode}_m(t), R_{\mathrm{str}}(c))\}$ for model m. Matching accepts exact, whitespace-normalized, or prefix matches, with explicit rejections taking precedence. The implemented EU proxy estimates probability mass outside $A_m(c)$; the AU proxy uses $\log |R_{\mathrm{str}}(c)|$ to measure diversity in the shared annotation space.

B.3 Annotation quality and coverage

The reasonable-token rate is 43.28% among seed candidates and 46.75% among proposed additions. The additions contribute 25.62% of the final reasonable tokens, showing that candidate expansion recovers alternatives beyond the source models' high-probability predictions. These are candidate-level coverage statistics, not probabilities under an evaluated LLM.

Against manually constructed sets at 200 continuation positions, the automatically constructed sets achieve 96% precision, 91% recall, and an F1 score of 93.4%. For a predicted set R and a manual set R^*, set precision and recall are based on $|R \cap R^*|/|R|$ and $|R \cap R^*|/|R^*|$, respectively. This check evaluates the annotated reasonable sets, whereas estimator correlations evaluate how well model-derived scores predict the resulting labels.

B.4 Audited benchmark examples

Figures 11–13 present three real benchmark positions: two English examples and one Chinese example. They cover a specific geological name and contexts admitting varied, semantically compatible descriptions. We inspect each complete conditioning prefix, recheck the stored annotations, and review pooled model predictions. The figures list all 3, 140, and 165 reasonable strings in the respective audited sets, together with a concrete continuation for each token. Tokens admitting the same continuation are grouped together; each listed token begins the text following the arrow. Whitespace-equivalent forms are merged, and large sets continue across pages without omitting entries.

The first English example completes the name Elkmont Sandstone in the ascending stratigraphic sequence of the Great Smoky Group. The Chinese example describes an in-progress course spanning academic terms, while the second English example describes a homemade sports drink whose ingredients and purpose are specified in the preceding context. A candidate may end inside a word: for example, `o` and `on` can both continue the prefix `Elkm` into `Elkmont`, just as `for` can begin `formula`. The displayed continuations illustrate why candidates are reasonable; they do not prescribe a unique subsequent wording. Full conditioning prefixes, source offsets, labels, and continuation witnesses accompany the case records.

English · Geological terminology

3 tokens

`wikipedia_english_06663 @ 2236`

The Great Smoky Group is reported as more than ~7,600 m (25,000 ft) thick in the park region and is commonly divided (ascending) into the Elkm |

Reasonable token(s) followed by one valid continuation; read down each column.

`o` → ont Sandstone, Thunderhead Sandstone, and Anakeesta Formation.	
`on` → ont Sandstone, Thunderhead Sandstone, and Anakeesta Formation.	
`ont` → ont Sandstone, Thunderhead Sandstone, and Anakeesta Formation.	

| marks the next-token boundary. Each listed token begins its paired continuation.

Figure 11: Geological terminology: all 3 reasonable tokens and their continuations.

Chinese · Course status

140 tokens · 1/2

`github_javascript_04754 @ 669`

```
const SPECIAL_TEXT = {
    'P': '通过', 'NP': '未通过', 'EX': '免修',
    'IP': '跨学期 |
```

Reasonable token(s) followed by one valid continuation; read down each column.

`' '}` → '};	`不` `不计` → 不计最终成绩（课程尚未完成）'
`', ','` → ','I': '缓考', 'W': '退课'};	`不可` → 不可确定最终成绩（课程仍在进行）'
`( (I (In` → (In Progress)'	`不能` → 不能确定最终成绩（课程仍在进行）'
`(P` → (Pending course completion)'	`仍` `仍在` → 仍在修读'
`(a` → (across semesters, course in progress)'	`仍未` → 仍未结课'
`(in` → (in progress)'	`保` `保持` → 保持进行中状态'
`,` → , 课程仍在进行'	`保留` → 保留为进行中状态'
`-` → -进行中'	`修` `修读` → 修读中'
`.` → .'	`修习` → 修习中'
`/` `/I` `/IP` → /IP'	`在` `在读` → 在读课程'
`/in` → /in progress'	`在学` → 在学课程'
`:` → : 课程仍在进行'	`在完成` → 在完成过程中'
`[` → [进行中]'	`在进行` → 在进行的课程'
`\` `\n` → \n（课程进行中）'	`学` `学习` `学习中` → 学习中'
`\r` → \r（课程进行中）'	`学分` → 学分待课程完成后结算'
`\u` → \u8bfe程进行中'	`完` `完成` → 完成前暂不记分'
`、` → 、进行中'	`尚` `尚未` → 尚未完成'
`。` → 。'	`已` → 已开始但尚未完成'
`上` `上课` → 上课中'	`延` `延续` → 延续课程，尚未结课'

| marks the next-token boundary. Each listed token begins its paired continuation.

Figure 12: Course status: all 140 reasonable tokens and their continuations.

Chinese · Course status **140 tokens · 2/2**

`github_javascript_04754 @ 669`

'IP': '跨学期 |

Reasonable token(s) followed by one valid continuation; read down each column.

开 开课 → 开课中'	续 → 续读中'
开始 → 开始修读，尚未结课'	缺 → 缺最终成绩（课程尚未完成）'
开设 → 开设的课程，尚未结束'	考 考核 → 考核尚未完成'
待 → 待结课'	考查 → 考查尚未完成'
待定 → 待定成绩（课程仍在进行）'	考评 → 考评尚未完成'
成 成绩 → 成绩尚未确定'	課 課程 → 課程仍在進行'
或 → 或跨学年，课程尚未完成'	计 计分 → 计分待结课'
或者 → 或者跨学年，课程尚未完成'	计入 → 计入绩点前需完成课程'
持 持续 → 持续修读中'	计算 → 计算最终成绩需待结课'
授 授课 → 授课中'	记 记分 → 记分待结课'
教 教学 教学中 → 教学中'	记录 → 记录为进行中'
无 → 无最终成绩（课程尚未完成）'	记载 → 记载为进行中'
暂 → 暂未结课'	评 评价 → 评价尚未完成'
暂时 → 暂时未结课'	评估 → 评估尚未完成'
有 → 有待完成'	评分 → 评分尚未完成'
期 期末 → 期末成绩尚未确定'	评定 → 评定尚未完成'
未 未修完 → 未修完'	课 课程 → 课程尚未完成'
未完 未完成 → 未完成'	课程的 → 课程的成绩尚未确定'
未定 → 未定成绩（课程仍在进行）'	还未 → 还未结课'
未有 → 未有最终成绩'	进 进程 进程中 → 进程中'
未结 → 未结课'	进度 → 进度：课程进行中'
正 正在 正在进行 → 正在进行'	进行 进行中 → 进行中'
状态 → 状态：进行中'	进行时 → 进行时'
的 的课程 → 的课程尚未结束'	进行的 → 进行的课程'
科 科目 → 科目进行中'	连 连续 → 连续修读中'
结 → 结课前暂不记分'	需 → 需继续修读'
结束 → 结束后结算成绩'	（ → （课程仍在进行）'
结算 → 结算成绩需待课程结束'	， → ，课程仍在进行'
继续 → 继续修读'	，成绩 → ，成绩尚未确定'
继续进行 → 继续进行'	，请 → ，请待课程结束后查询最终成绩'
	： → ：课程仍在进行'

| marks the next-token boundary. Each listed token begins its paired continuation.

Figure 12: Chinese: Course status (continued).

English · Sports-drink formulation

165 tokens · 1/2

`bbc_news_03494 @ 2225`

"Combine two-thirds water, one-third fruit juice like pineapple juice and a pinch of salt until you can just about taste it.
"That gives you a well-designed |

Reasonable token(s) followed by one valid continuation; read down each column.

- `'` → 'sports drink' made from ordinary ingredients.
- `(` → (and inexpensive) sports drink.
- `1` → 1:2 juice-to-water mixture with added salt.
- `2` → 2:1 water-to-juice mixture with added salt.
- `3` → 3-ingredient sports drink.
- `6` → 6% carbohydrate and electrolyte solution.
- `DIY` → DIY sports drink.
- `Na` → NaCl-containing sports drink.
- `a` `al` `all` → all-natural sports drink.
- `aid` → aid to hydration during prolonged exercise.
- `alternative` → alternative to an expensive sports drink.
- `an` `and` → and inexpensive sports drink.
- `approx` `approximately` → approximately 6% carbohydrate and electrolyte solution.
- `around` → around 6% carbohydrate and electrolyte solution.
- `at` → athlete-friendly sports drink.
- `b` `basic` → basic sports drink.
- `balance` `balanced` → balanced mixture of water, juice and salt.
- `base` → base for a homemade sports drink.
- `be` `beverage` → beverage for prolonged exercise.
- `blend` → blend of water, fruit juice and salt.
- `body` → body-fluid replacement drink.
- `but` → but inexpensive sports drink.
- `calorie` → calorie-containing drink.
- `car` `carb` `carbs` → carbs-and-electrolytes drink.
- `carbohydrate` → carbohydrate and electrolyte drink.
- `cheap` → cheap alternative to commercial sports drinks.
- `choice` → choice for hydration during long exercise sessions.
- `cocktail` → cocktail of water, juice and salt.
- `combination` → combination of fluid, carbohydrate and sodium.
- `commercial` → commercial-style sports drink made at home.
- `con` `conco` → concoction for hydration during prolonged exercise.
- `cost` → cost-effective sports drink.
- `d` `dr` `drink` → drink for prolonged exercise.
- `de` → dehydration-prevention drink for use during prolonged exercise.
- `diet` → dietary source of fluid and carbohydrate.
- `dil` `diluted` → diluted fruit-juice drink with added salt.
- `drinking` → drinking solution for prolonged exercise.
- `e` `ex` `exercise` → exercise drink.
- `easily` → easily prepared sports drink.
- `easy` → easy-to-make sports drink.
- `effective` → effective way to replace fluid during prolonged exercise.
- `elect` `electro` `electroly` `electrolyte` → electrolyte-containing drink.
- `energy` → energy drink containing carbohydrate.
- `f` `fit` → fitness drink.
- `fl` `flavor` → flavorful sports drink.
- `flavour` → flavourful sports drink.
- `flu` `fluid` → fluid-replacement drink.
- `for` `form` `formula` → formula for a homemade sports drink.
- `formulation` → formulation for use during prolonged exercise.
- `from` → from-scratch sports drink.
- `fruit` → fruit-juice-based sports drink.
- `fu` `functional` → functional sports drink.
- `fuel` → fuel source for prolonged exercise.
- `g` `glucose` → glucose-containing sports drink.
- `good` → good source of fluid and carbohydrate.
- `gym` → gym-workout drink for prolonged sessions.
- `h` `hy` `hyd` `hydr` `hydration` → hydration drink.
- `high` → high-water-content sports drink.

| marks the next-token boundary. Each listed token begins its paired continuation.

Figure 13: Sports-drink formulation: all 165 reasonable tokens and their continuations.

English · Sports-drink formulation **165 tokens · 2/2**

`bbc_news_03494 @ 2225`

"That gives you a well-designed |

Reasonable token(s) followed by one valid continuation; read down each column.

`hom` `home` `homem` `homemade` → homemade sports drink.
`i` `in` → inexpensive sports drink.
`ion` `ionic` → ionic solution containing sodium and carbohydrate.
`j` `juice` → juice-based sports drink.
`kind` → kind of sports drink.
`less` → less expensive alternative to a powdered sports drink.
`liquid` → liquid source of carbohydrate and sodium.
`low` → low-cost sports drink.
`m` `mineral` → mineral-containing drink.
`mix` → mix for prolonged exercise.
`mixed` → mixed drink containing juice and salt.
`mixture` → mixture of water, juice and salt.
`muscle` → muscle-fuelling drink for prolonged exercise.
`natural` `naturally` → naturally flavoured sports drink.
`non` → non-commercial sports drink.
`nutrient` → nutrient-containing drink.
`nutrition` `nutritional` → nutritional drink for prolonged exercise.
`on` → on-the-go sports drink.
`one` → one-part-juice, two-parts-water sports drink with added salt.
`or` `oral` → oral rehydration drink for use during exercise.
`p` `potassium` → potassium-containing drink.
`per` `performance` → performance-supporting drink for prolonged exercise.
`post` → post-exercise drink to replace fluid and carbohydrate.
`pre` → pre-exercise hydration drink.
`pretty` → pretty inexpensive sports drink.
`product` → product you can make at home.
`quick` → quick-to-prepare sports drink.
`ratio` → ratio of water to juice, with added salt.
`re` `refresh` `refreshing` → refreshing sports drink.
`ready` → ready-to-drink sports beverage.
`recipe` → recipe for a homemade sports drink.
`recovery` → recovery drink for replacing fluid and carbohydrate after exercise.
`reh` `rehyd` → rehydration drink for use during exercise.
`replacement` → replacement for an expensive sports drink.
`replen` → replenishment drink for fluid and carbohydrate.
`roughly` → roughly 6% carbohydrate and electrolyte solution.
`s` `sport` → sport drink.
`sal` `salt` → salt-containing sports drink.
`salty` → salty drink containing water and carbohydrate.
`self` → self-made sports drink.
`simple` → simple sports drink.
`six` → six per cent carbohydrate and electrolyte solution.
`slightly` → slightly salty sports drink.
`sodium` → sodium-containing drink.
`solution` → solution containing carbohydrate and salt.
`source` → source of fluid and carbohydrate.
`sporting` → sporting-event hydration drink.
`sports` → sports drink.
`sugar` → sugar-containing sports drink.
`supplement` → supplement to water during prolonged exercise.
`sweat` → sweat-replacement drink for prolonged exercise.
`sweet` → sweet-and-salty sports drink.
`t` `tasty` → tasty sports drink.
`th` `thirst` → thirst-quenching sports drink.
`v` `very` → very inexpensive sports drink.
`water` → water-based sports drink.
`well` → well-balanced mixture of water, juice and salt.
`workout` → workout drink.

| marks the next-token boundary. Each listed token begins its paired continuation.

Figure 13: English: Sports-drink formulation (continued).

C Evaluation Protocols and Complete Results

We specify the scoring and aggregation protocols for Section 3, followed by complete model and estimator results.

C.1 LLM evaluation

Aggregation and comparison. The continuation evaluation covers 67 LLMs on the 7,216-position benchmark. The overall unreasonable continuation rate weights positions equally; the macro rate first averages within each domain and then gives the 15 domains equal weight. Table 3 reports both summaries. Tables 4–6 give all domain-level results in the same model order. Perplexity, bits per byte, and Artificial Analysis scores are compared only where the corresponding readout is available. When both thinking and non-thinking Artificial Analysis scores are reported, the comparison uses the thinking score.

Table 3: Complete model ranking. Overall and macro unreasonable continuation rates are percentages (lower is better); mean EU is on the probability scale.

Rank	Model	Overall	Mean EU	Macro
1	Kimi-K3	9.78	0.0978	9.55
2	DeepSeek-V4-Pro-0813	9.94	0.0994	9.59
3	Qwen3.8-2.4T-A95B	10.46	0.1046	10.04
4	DeepSeek-V4-Pro-Base	11.02	0.1102	10.66
5	DeepSeek-V4-Flash	11.43	0.1143	10.90
6	Qwen3.5-397B-A17B	11.51	0.1151	11.40
7	Kimi-K2.5	11.91	0.1191	11.65
8	MiMo-V2.5-Pro-Base	12.23	0.1223	11.86
9	GLM-5.3	12.27	0.1227	11.83
10	DeepSeek-V3.2	12.48	0.1248	12.22
11	MiniMax-M3	12.53	0.1253	12.02
12	DeepSeek-V4-Flash-Base	12.61	0.1261	12.25
13	Hy3	12.99	0.1299	12.72
14	MiMo-V2.5-Pro	13.01	0.1301	12.43
15	GLM-4.5	13.19	0.1319	12.90
16	DeepSeek-V3	13.27	0.1327	12.80
17	Kimi-K2-Instruct	13.57	0.1357	13.46
18	Kimi-K2-Base	13.59	0.1359	13.46
19	Qwen3.8-Flash-Next	13.79	0.1379	13.58
20	Hy3-preview	13.93	0.1393	13.62
21	MiMo-V2.5-Base	14.04	0.1404	13.66
22	GLM-4.7	14.13	0.1413	13.68
23	Hy3-preview-Base	14.41	0.1441	14.13
24	GLM-4.5-Base	14.42	0.1442	14.12
25	DeepSeek-V3-Base	14.53	0.1453	14.13
26	Mistral-Small-4-119B-2603	14.98	0.1498	14.83
27	Qwen3-235B-A22B-Thinking-2507	15.34	0.1534	15.23
28	Step-3.5-Flash-Base	15.68	0.1568	15.28
29	Qwen3-235B-A22B	16.09	0.1609	15.93
30	Step-3.5-Flash-nothink	16.14	0.1614	15.76
31	Step-3.5-Flash	16.20	0.1620	15.84
32	Qwen3.5-122B-A10B	16.54	0.1654	16.63
33	MiMo-V2.5	17.37	0.1737	16.54
34	GLM-4.5-Air	18.19	0.1819	17.61
35	GLM-4.5-Air-Base	18.49	0.1849	17.97
36	Qwen3-30B-A3B	18.60	0.1860	18.57
37	Seed-OSS-36B-Instruct	19.21	0.1921	19.34
38	Qwen3.6-35B-A3B	19.43	0.1943	19.39
39	Qwen3.5-27B	19.46	0.1946	19.59
40	Qwen3-14B	19.71	0.1971	19.47
41	Qwen3.5-35B-A3B-Base	19.89	0.1989	19.95
42	Qwen3.8-27B	19.93	0.1993	19.74
43	Qwen3-32B	20.03	0.2003	20.26

Rank	Model	Overall	Mean EU	Macro
44	Qwen3.5-35B-A3B	20.57	0.2057	20.70
45	Llama-4-Scout-17B-16E-Instruct	20.59	0.2059	20.11
46	Mistral-Small-24B-Base-2501	20.71	0.2071	20.21
47	Qwen3-8B	21.39	0.2139	21.22
48	Qwen3-30B-A3B-Base	21.66	0.2166	21.91
49	Seed-OSS-36B-Base	21.91	0.2191	22.04
50	Mistral-Small-24B-Instruct-2501	22.22	0.2222	21.77
51	Qwen3-14B-Base	22.79	0.2279	22.79
52	Qwen3.5-9B-Base	24.15	0.2415	24.45
53	Qwen3-4B	24.32	0.2432	24.43
54	Qwen3.5-9B	24.93	0.2493	25.15
55	Qwen3-8B-Base	25.32	0.2532	25.48
56	Qwen3-1.7B	27.16	0.2716	27.55
57	Qwen3.5-4B-Base	27.55	0.2755	28.00
58	Qwen3.5-4B	28.28	0.2828	28.65
59	Qwen3-4B-Base	28.39	0.2839	28.84
60	gpt-oss-120b	28.47	0.2847	33.61
61	Qwen3-1.7B-Base	31.96	0.3196	32.58
62	Qwen3.5-2B-Base	32.63	0.3263	33.17
63	Qwen3.5-2B	33.14	0.3314	33.69
64	Qwen3-0.6B	35.28	0.3528	36.13
65	Qwen3-0.6B-Base	36.57	0.3657	37.39
66	Qwen3.5-0.8B-Base	36.64	0.3664	37.58
67	Qwen3.5-0.8B	36.71	0.3671	37.71

Table 4: Unreasonable continuation rates on research domains (%; lower is better). Models follow the overall ranking. Other denotes the remaining research subjects.

Model	CS	Math	Other	Physics	Biology
Kimi-K3	11.5	9.6	11.1	11.4	18.4
DeepSeek-V4-Pro-0813	10.7	9.9	11.2	11.8	17.4
Qwen3.8-2.4T-A95B	12.3	10.4	12.3	13.5	19.4
DeepSeek-V4-Pro-Base	12.6	10.7	12.1	13.0	19.2
DeepSeek-V4-Flash	13.0	11.6	12.9	13.1	19.0
Qwen3.5-397B-A17B	12.5	9.5	11.2	13.3	20.0
Kimi-K2.5	13.3	12.9	12.8	14.4	20.7
MiMo-V2.5-Pro-Base	13.2	14.0	14.5	14.3	21.3
GLM-5.3	14.1	14.1	13.9	14.1	21.2
DeepSeek-V3.2	14.7	12.3	13.4	13.4	20.7
MiniMax-M3	14.1	16.1	14.0	14.0	20.1
DeepSeek-V4-Flash-Base	14.2	12.9	13.8	14.4	21.1
Hy3	13.0	14.4	14.7	14.8	23.4
MiMo-V2.5-Pro	14.4	15.7	16.1	15.0	22.3
GLM-4.5	14.8	15.3	14.0	15.1	20.5
DeepSeek-V3	15.8	13.3	14.8	14.3	21.6
Kimi-K2-Instruct	16.4	14.1	14.2	15.0	22.7
Kimi-K2-Base	15.8	14.2	14.4	15.3	21.4
Qwen3.8-Flash-Next	14.0	13.2	15.4	17.0	25.5
Hy3-preview	14.4	15.1	15.5	15.8	25.2
MiMo-V2.5-Base	15.1	16.2	15.2	17.6	23.8
GLM-4.7	17.1	16.1	14.5	17.0	23.4
Hy3-preview-Base	14.5	15.9	15.5	17.2	25.7
GLM-4.5-Base	16.5	15.8	15.5	16.5	22.6
DeepSeek-V3-Base	16.6	14.7	16.0	15.2	23.0
Mistral-Small-4-119B-2603	16.1	16.4	16.3	16.6	21.3
Qwen3-235B-A22B-Thinking-2507	17.4	13.5	16.6	17.1	25.1
Step-3.5-Flash-Base	17.1	18.0	17.5	18.1	24.4
Qwen3-235B-A22B	17.5	15.0	17.5	17.3	25.6
Step-3.5-Flash-nothink	18.3	18.3	17.9	17.9	24.0

Model	CS	Math	Other	Physics	Biology
Step-3.5-Flash	18.4	18.4	17.8	17.9	24.2
Qwen3.5-122B-A10B	19.0	13.4	16.7	17.9	27.6
MiMo-V2.5	22.9	20.8	19.8	22.5	28.3
GLM-4.5-Air	21.9	21.5	20.0	20.8	27.3
GLM-4.5-Air-Base	21.2	22.0	20.6	21.0	28.2
Qwen3-30B-A3B	21.1	17.1	19.5	21.5	30.4
Seed-OSS-36B-Instruct	20.6	19.9	18.2	21.3	28.3
Qwen3.6-35B-A3B	21.5	18.1	20.5	22.3	33.5
Qwen3.5-27B	22.0	15.6	19.6	20.0	32.5
Qwen3-14B	20.6	19.2	20.6	21.0	31.8
Qwen3.5-35B-A3B-Base	21.0	17.2	19.4	21.3	32.7
Qwen3.8-27B	19.2	18.4	20.4	23.5	35.2
Qwen3-32B	21.3	16.2	18.6	21.6	31.9
Qwen3.5-35B-A3B	23.2	18.0	20.5	21.4	33.8
Llama-4-Scout-17B-16E-Instruct	22.6	21.5	21.3	22.0	33.0
Mistral-Small-24B-Base-2501	22.0	22.6	20.9	21.0	29.1
Qwen3-8B	20.5	20.6	20.9	24.3	34.3
Qwen3-30B-A3B-Base	23.6	18.5	23.0	23.4	33.5
Seed-OSS-36B-Base	22.6	22.1	22.1	25.2	32.1
Mistral-Small-24B-Instruct-2501	23.3	24.7	22.6	22.7	30.4
Qwen3-14B-Base	24.2	20.9	23.4	24.4	36.2
Qwen3.5-9B-Base	24.0	21.7	23.1	24.2	37.7
Qwen3-4B	25.7	23.3	24.6	26.1	35.8
Qwen3.5-9B	26.0	23.2	24.1	25.1	38.5
Qwen3-8B-Base	26.0	23.7	26.3	27.7	38.3
Qwen3-1.7B	27.3	26.5	26.1	28.1	40.3
Qwen3.5-4B-Base	27.7	24.7	26.3	28.2	40.3
Qwen3.5-4B	30.2	25.1	27.3	29.0	41.7
Qwen3-4B-Base	29.3	25.8	29.1	30.1	41.9
gpt-oss-120b	28.9	28.3	25.9	26.2	35.2
Qwen3-1.7B-Base	32.2	29.8	31.7	33.0	45.4
Qwen3.5-2B-Base	33.4	29.6	30.5	33.4	45.5
Qwen3.5-2B	33.4	30.7	30.9	33.2	46.4
Qwen3-0.6B	35.1	33.4	33.2	35.2	48.5
Qwen3-0.6B-Base	37.0	34.3	34.4	37.2	49.5
Qwen3.5-0.8B-Base	36.2	32.8	34.1	36.5	48.3
Qwen3.5-0.8B	36.1	33.7	33.9	35.9	48.2

Table 5: Unreasonable continuation rates on code domains (%; lower is better). Models follow the overall ranking. Other denotes the remaining programming languages.

Model	C++	JavaScript	Markdown	Other	Python
Kimi-K3	5.1	4.9	8.4	7.5	6.5
DeepSeek-V4-Pro-0813	6.5	5.9	9.0	8.1	9.1
Qwen3.8-2.4T-A95B	5.6	5.7	9.0	7.2	8.5
DeepSeek-V4-Pro-Base	6.6	6.5	11.4	9.4	9.3
DeepSeek-V4-Flash	7.7	7.2	13.1	9.4	9.5
Qwen3.5-397B-A17B	7.4	6.7	11.1	9.3	11.0
Kimi-K2.5	5.9	5.9	10.4	8.8	8.9
MiMo-V2.5-Pro-Base	6.1	6.6	11.7	8.3	8.9
GLM-5.3	6.9	6.4	10.6	9.6	9.8
DeepSeek-V3.2	7.4	7.2	12.9	10.5	10.6
MiniMax-M3	7.8	7.0	11.0	9.0	8.0
DeepSeek-V4-Flash-Base	8.1	7.6	13.4	9.9	10.4
Hy3	7.6	7.0	10.8	8.2	9.1
MiMo-V2.5-Pro	6.5	6.7	12.4	9.9	9.2
GLM-4.5	7.5	7.5	12.5	10.7	10.4
DeepSeek-V3	8.5	8.2	14.5	11.3	11.3
Kimi-K2-Instruct	7.2	7.0	11.9	9.7	9.8

Model	C++	JavaScript	Markdown	Other	Python
Kimi-K2-Base	7.5	7.7	12.0	10.6	10.3
Qwen3.8-Flash-Next	7.7	7.0	11.6	10.2	9.6
Hy3-preview	8.3	7.9	12.1	8.0	9.8
MiMo-V2.5-Base	7.2	6.9	11.9	9.5	10.0
GLM-4.7	7.5	8.0	12.7	11.0	10.3
Hy3-preview-Base	8.3	7.4	12.9	9.1	9.8
GLM-4.5-Base	8.4	7.8	14.6	11.4	11.3
DeepSeek-V3-Base	9.2	9.5	16.2	12.2	12.5
Mistral-Small-4-119B-2603	9.9	9.7	14.9	12.7	12.4
Qwen3-235B-A22B-Thinking-2507	9.6	9.0	15.3	11.8	10.5
Step-3.5-Flash-Base	8.8	9.0	16.1	11.9	10.1
Qwen3-235B-A22B	10.5	10.5	16.0	12.9	12.2
Step-3.5-Flash-nothink	10.0	8.7	16.7	12.6	11.9
Step-3.5-Flash	10.0	8.8	16.7	12.8	11.9
Qwen3.5-122B-A10B	11.2	9.7	16.1	13.2	12.6
MiMo-V2.5	8.6	8.4	13.6	12.6	10.9
GLM-4.5-Air	11.6	9.5	16.9	13.8	13.3
GLM-4.5-Air-Base	11.1	10.0	18.8	14.2	13.4
Qwen3-30B-A3B	10.3	10.5	18.9	13.4	12.3
Seed-OSS-36B-Instruct	12.5	9.1	15.6	15.0	13.1
Qwen3.6-35B-A3B	12.3	10.3	17.3	14.6	13.3
Qwen3.5-27B	14.5	11.0	19.1	16.2	14.2
Qwen3-14B	13.0	11.9	19.8	15.0	14.2
Qwen3.5-35B-A3B-Base	14.5	11.5	20.2	17.3	15.4
Qwen3.8-27B	14.5	10.3	15.6	14.4	13.8
Qwen3-32B	12.7	13.3	21.6	16.6	15.1
Qwen3.5-35B-A3B	14.9	11.7	21.3	16.7	14.8
Llama-4-Scout-17B-16E-Instruct	14.2	14.1	21.0	17.0	16.8
Mistral-Small-24B-Base-2501	15.8	14.9	24.0	18.4	17.8
Qwen3-8B	15.4	13.1	21.9	16.6	14.7
Qwen3-30B-A3B-Base	13.8	12.4	23.4	16.7	14.7
Seed-OSS-36B-Base	14.3	10.5	19.0	16.3	13.7
Mistral-Small-24B-Instruct-2501	16.8	15.5	24.7	20.0	18.2
Qwen3-14B-Base	15.5	13.5	22.7	17.9	15.5
Qwen3.5-9B-Base	18.5	16.5	23.7	20.3	19.4
Qwen3-4B	16.5	15.1	24.2	18.5	17.8
Qwen3.5-9B	18.7	16.4	25.0	20.9	19.4
Qwen3-8B-Base	17.4	14.6	25.3	20.5	18.1
Qwen3-1.7B	18.5	17.6	26.0	22.4	21.3
Qwen3.5-4B-Base	21.3	18.4	26.6	23.5	22.2
Qwen3.5-4B	21.7	18.9	26.8	23.8	22.3
Qwen3-4B-Base	19.0	17.2	27.8	23.2	20.7
gpt-oss-120b	11.3	9.8	19.0	13.7	14.5
Qwen3-1.7B-Base	22.5	20.9	30.9	27.1	25.1
Qwen3.5-2B-Base	26.4	22.9	32.6	28.3	25.8
Qwen3.5-2B	26.9	22.7	33.0	29.0	26.3
Qwen3-0.6B	26.6	24.2	35.0	31.3	27.5
Qwen3-0.6B-Base	27.4	25.9	37.1	31.7	29.3
Qwen3.5-0.8B-Base	30.5	26.5	37.7	31.9	31.1
Qwen3.5-0.8B	30.5	26.2	37.3	32.7	30.9

Table 6: Unreasonable continuation rates on news, encyclopedia, and fiction domains (%; lower is better). Models follow the overall ranking. En denotes English and Other denotes the remaining subjects or languages within the source family.

Model	BBC	Wiki-En	Wiki-Other	AO3-En	AO3-Other
Kimi-K3	5.3	8.3	9.1	13.3	12.8
DeepSeek-V4-Pro-0813	5.9	8.3	7.2	11.8	11.3
Qwen3.8-2.4T-A95B	6.2	9.2	7.1	14.2	10.1

Model	BBC	Wiki-En	Wiki-Other	AO3-En	AO3-Other
DeepSeek-V4-Pro-Base	6.6	9.7	7.9	13.0	11.9
DeepSeek-V4-Flash	5.8	9.8	8.7	11.9	10.6
Qwen3.5-397B-A17B	7.6	11.7	12.7	14.7	12.5
Kimi-K2.5	6.9	9.0	13.7	14.7	16.4
MiMo-V2.5-Pro-Base	6.5	10.4	12.1	14.1	15.8
GLM-5.3	6.1	10.2	10.4	14.7	15.3
DeepSeek-V3.2	7.2	10.1	13.2	14.1	15.7
MiniMax-M3	6.7	11.3	13.8	13.7	13.7
DeepSeek-V4-Flash-Base	7.6	11.2	10.3	14.2	14.7
Hy3	7.2	12.1	14.8	15.3	18.4
MiMo-V2.5-Pro	5.9	9.5	12.1	15.1	15.6
GLM-4.5	7.9	11.9	13.4	14.3	17.7
DeepSeek-V3	6.7	9.5	12.8	14.2	15.0
Kimi-K2-Instruct	8.7	12.3	18.0	14.6	20.1
Kimi-K2-Base	8.3	11.6	17.7	15.4	19.6
Qwen3.8-Flash-Next	11.6	13.7	14.7	18.2	14.2
Hy3-preview	7.8	12.4	16.8	16.5	18.6
MiMo-V2.5-Base	7.4	13.4	16.3	16.6	18.0
GLM-4.7	7.8	11.9	14.8	14.0	19.4
Hy3-preview-Base	8.7	13.9	16.8	16.6	19.9
GLM-4.5-Base	8.1	12.0	15.3	14.9	21.0
DeepSeek-V3-Base	7.8	11.8	15.5	14.9	16.8
Mistral-Small-4-119B-2603	10.5	13.8	15.1	18.4	18.5
Qwen3-235B-A22B-Thinking-2507	9.4	15.5	20.1	19.4	18.2
Step-3.5-Flash-Base	9.4	13.3	19.8	17.0	18.7
Qwen3-235B-A22B	8.7	16.8	19.5	20.9	18.1
Step-3.5-Flash-nothink	10.7	14.4	20.0	17.3	17.6
Step-3.5-Flash	10.9	14.5	20.4	17.1	17.9
Qwen3.5-122B-A10B	13.1	17.1	20.3	21.0	20.6
MiMo-V2.5	10.2	13.4	18.1	17.0	20.9
GLM-4.5-Air	8.7	15.8	19.6	18.3	25.1
GLM-4.5-Air-Base	9.7	16.0	18.4	19.4	25.4
Qwen3-30B-A3B	13.1	20.4	26.4	22.6	21.2
Seed-OSS-36B-Instruct	12.6	19.0	34.7	16.6	33.5
Qwen3.6-35B-A3B	16.2	20.9	21.7	24.2	24.0
Qwen3.5-27B	15.7	21.5	24.8	22.7	24.6
Qwen3-14B	11.6	20.8	29.1	21.0	22.6
Qwen3.5-35B-A3B-Base	15.4	21.5	22.6	23.6	25.5
Qwen3.8-27B	15.1	23.4	26.1	24.3	21.9
Qwen3-32B	14.2	24.0	29.5	24.8	22.7
Qwen3.5-35B-A3B	17.3	22.1	23.6	25.1	26.0
Llama-4-Scout-17B-16E-Instruct	13.5	20.8	19.8	23.8	20.3
Mistral-Small-24B-Base-2501	11.7	18.8	22.6	20.1	23.6
Qwen3-8B	16.1	20.9	30.5	24.0	24.6
Qwen3-30B-A3B-Base	16.5	24.7	32.0	24.7	27.7
Seed-OSS-36B-Base	14.0	22.8	38.7	19.5	37.5
Mistral-Small-24B-Instruct-2501	13.9	20.5	26.1	22.2	25.0
Qwen3-14B-Base	16.5	25.7	33.5	24.6	27.5
Qwen3.5-9B-Base	21.3	27.0	31.4	27.3	30.6
Qwen3-4B	16.9	27.0	37.8	24.8	32.3
Qwen3.5-9B	22.0	27.2	31.6	28.2	30.7
Qwen3-8B-Base	21.9	26.9	36.3	27.9	31.3
Qwen3-1.7B	23.7	30.7	40.2	30.1	34.4
Qwen3.5-4B-Base	22.9	33.1	37.2	30.6	36.9
Qwen3.5-4B	24.0	33.0	38.3	31.0	36.6
Qwen3-4B-Base	23.4	33.2	43.4	31.1	37.4
gpt-oss-120b	56.7	56.3	57.2	61.9	59.3
Qwen3-1.7B-Base	29.1	37.7	47.6	34.9	40.8
Qwen3.5-2B-Base	29.1	36.9	44.6	36.4	42.5
Qwen3.5-2B	29.9	37.2	46.3	36.2	43.2
Qwen3-0.6B	34.1	41.1	51.3	38.1	47.4
Qwen3-0.6B-Base	34.0	43.2	52.8	39.6	47.5

Model	BBC	Wiki-En	Wiki-Other	AO3-En	AO3-Other
Qwen3.5-0.8B-Base	32.8	42.0	51.0	42.0	50.5
Qwen3.5-0.8B	34.2	42.0	51.2	41.7	51.2

C.2 Uncertainty estimators and comparison protocol

Four estimator families. The 35 configurations comprise eight probability-based estimators, 11 based on model internals, 15 Bayesian readouts, and one generative confidence estimator. Tables 7 and 8 give the complete results and naming conventions. A configuration denotes one reported readout, so TU, AU, and EU from the same sampling method count as separate configurations. The internal-state family uses the prefix I, keeping it distinct from entropy H.

Probability readouts. Write $p_t = P(t \mid c)$, $V = |\mathcal{V}|$, and let w denote the reference continuation token. The basic readouts are Shannon entropy $-\sum_t p_t \log p_t$, MaxProb $p_{(1)}$, min-entropy $-\log p_{(1)}$, probability margin $p_{(1)} - p_{(2)}$, and reference-token probability (TokenProb) p_w. Self-Certainty (Kang et al., 2025) is

$$\mathrm{SC}(P) = D_{\mathrm{KL}}(U\|P) = -\log V - \frac{1}{V}\sum_t \log p_t, \tag{10}$$

where U is uniform over the vocabulary. Its negative is the uncertainty-oriented score. For a fixed model, its absolute Spearman correlations equal those of mean log probability. Min-K%++ (Zhang et al., 2024) uses a standardized reference-token log probability: $(\log p_w - \mu)/\sigma$, where $\mu = \sum_t p_t \log p_t$ and $\sigma^2 = \sum_t p_t(\log p_t - \mu)^2$. At a single annotated position this is a token readout; sequence-level selection of the lowest-scoring fraction is a separate aggregation step.

Internal and Bayesian readouts. Internal estimators use logits, intermediate-layer predictions, hidden representations, or auxiliary prediction heads. For example, negative energy depends on the unnormalized log-sum-exp of logits, while DoLa-style readouts (Chuang et al., 2024b) summarize Jensen–Shannon divergence between layer predictions. Representation-based scores measure dispersion or distance in hidden-state space (Chen et al., 2024). Auxiliary-head methods are evaluated only on models exposing the necessary heads.

For sampled distributions $P^{(1)}, \ldots, P^{(S)}$, let $P_{\mathrm{avg}} = S^{-1}\sum_s P^{(s)}$. The entropy decomposition underlying MC dropout (Gal & Ghahramani, 2016) is

$$\mathrm{TU} = H(P_{\mathrm{avg}}), \qquad \mathrm{AU} = \frac{1}{S}\sum_s H(P^{(s)}), \qquad \mathrm{EU} = \mathrm{TU} - \mathrm{AU}. \tag{11}$$

Related readouts apply perturbations to representations or mixture-of-experts routing; evidence-based methods instead estimate concentration parameters (Zhang et al., 2025; Ma et al., 2025). The generative configuration queries the model for confidence (Kadavath et al., 2022). These families differ in required access: predictive probabilities, model internals, repeated forward passes, or additional generation.

Correlations and purity. For each supported model, Spearman correlation compares an estimator's scores with position-level EU and AU proxy labels. We aggregate the absolute within-model correlations separately for EU and AU, denoting the resulting magnitudes by $|\rho_{EU}|$ and $|\rho_{AU}|$.[3] Purity is computed from these unrounded values:

$$\pi_{EU} = \frac{|\rho_{EU}|}{|\rho_{EU}| + |\rho_{AU}|}. \tag{12}$$

Constant scores or labels yield undefined correlations and are excluded; purity is undefined if both aggregate magnitudes are zero. Absolute correlations assess signal strength without selecting a deployment direction, and high purity alone does not imply high EU correlation.

[3]Final aggregates of correlations, normalized ridge coefficients, and within-model logit differences use medians over the specified groups in this appendix and Appendix D.

Experiment coverage. We evaluate estimators on the 7,216-position UNCERTAINBENCH. Coverage depends on each method's architectural requirements and access to the required outputs. For example, MTP-based estimators require compatible multi-token prediction heads, while MoE perturbation requires a mixture-of-experts architecture with accessible routers.

Table 7: Complete estimator results. Correlation magnitudes are aggregated across models before computing EU purity. Each row is a method configuration evaluated where its required architecture and outputs are available.

Code	Configuration	$\lvert\rho_{EU}\rvert$	$\lvert\rho_{AU}\rvert$	π_{EU} (%)
P_{SelfCert}	Self-Certainty	0.4013	0.4167	49.06
P_{Shannon}	Shannon entropy	0.3847	0.5072	43.13
P_{MinEnt}	Min-entropy	0.3175	0.4299	42.48
P_{MaxProb}	Maximum probability	0.3175	0.4299	42.48
$P_{\text{MinK++}}$	Min-K%++	0.3044	0.5077	37.48
P_{Margin}	Probability margin	0.2433	0.3500	41.01
$P_{\text{TokenProb}}$	Reference-token probability	0.2115	0.3637	36.77
P_{DeepConf}	DeepConf	0.1404	0.2551	35.50
$I_{\text{NegEnergy}}$	Negative energy	0.3656	0.4147	46.85
$I_{\text{DoLa-max}}$	DoLa: maximum JSD	0.2488	0.3751	39.87
$I_{\text{MTP-JSD}}$	MTP: JSD	0.1626	0.2707	37.53
$I_{\text{Cos-JSD}}$	Hidden cosine distance $\times$ JSD	0.1626	0.2696	37.62
$I_{\text{MTP-agree}}$	MTP: top-1 agreement	0.1296	0.2125	37.87
$I_{\text{DoLa-mean}}$	DoLa: mean JSD	0.1027	0.0824	55.49
$I_{\text{DTR-depth}}$	Settling depth	0.0799	0.0542	59.60
$I_{\text{MTP-cos}}$	MTP: cosine distance	0.0575	0.0614	48.34
$I_{\text{DTR-deep}}$	Deep-settling indicator	0.0515	0.0438	54.01
$I_{\text{EigenScore}}$	EigenScore	0.0485	0.1737	21.84
$I_{\text{Mahalanobis}}$	Mahalanobis distance	0.0386	0.0722	34.85
$B_{\text{MC-TU}}$	MC dropout: TU	0.3846	0.5011	43.42
$B_{\text{MC-AU}}$	MC dropout: AU	0.3799	0.5146	42.47
$B_{\text{TokUR-TU}}$	TokUR: TU	0.3670	0.4877	42.94
$B_{\text{TokUR-AU}}$	TokUR: AU	0.3630	0.5061	41.77
$B_{\text{MoE-TU}}$	MoE perturbation: TU	0.3611	0.5249	40.75
$B_{\text{MoE-AU}}$	MoE perturbation: AU	0.3582	0.5305	40.31
$B_{\text{LogTokU-Sum}}$	LogTokU: concentration	0.3427	0.3148	52.13
$B_{\text{LogTokU-EU}}$	LogTokU: EU	0.3415	0.2981	53.39
$B_{\text{MC-Top1}}$	MC dropout: top-1 probability	0.3192	0.4148	43.49
$B_{\text{TokUR-Top1}}$	TokUR: top-1 probability	0.3094	0.3831	44.68
$B_{\text{MoE-Top1}}$	MoE perturbation: top-1 probability	0.3042	0.4469	40.50
$B_{\text{MC-EU}}$	MC dropout: EU	0.2648	0.0842	75.87
$B_{\text{TokUR-EU}}$	TokUR: EU	0.2314	0.1130	67.20
$B_{\text{MoE-EU}}$	MoE perturbation: EU	0.2187	0.2422	47.44
$B_{\text{LogTokU-AU}}$	LogTokU: AU	0.0979	0.1518	39.20
G_{SelfEval}	Self-evaluated confidence	0.0268	0.0323	45.36

Table 8: Estimator naming and readouts. Category prefixes are P (probability), I (model internals), B (Bayesian), and G (generative).

Code	Configuration	Readout
P_{SelfCert}	Self-Certainty	KL from a uniform vocabulary distribution to the prediction
P_{Shannon}	Shannon entropy	Predictive entropy
P_{MinEnt}	Min-entropy	Negative log maximum probability
P_{MaxProb}	Maximum probability	Largest next-token probability
$P_{\text{MinK++}}$	Min-K%++	Standardized reference-token log probability
P_{Margin}	Probability margin	Gap between the two largest probabilities
$P_{\text{TokenProb}}$	Reference-token probability	Requires the original continuation token
P_{DeepConf}	DeepConf	Token-level confidence readout
$I_{\text{NegEnergy}}$	Negative energy	Unnormalized logit statistic

Code	Configuration	Readout
$I_{\text{DoLa-max}}$	DoLa: maximum JSD	Maximum divergence across layer predictions
$I_{\text{MTP-JSD}}$	MTP: JSD	Divergence between prediction heads
$I_{\text{Cos-JSD}}$	Hidden cosine distance $\times$ JSD	Representation and prediction disagreement
$I_{\text{MTP-agree}}$	MTP: top-1 agreement	Agreement between prediction heads
$I_{\text{DoLa-mean}}$	DoLa: mean JSD	Mean divergence across layer predictions
$I_{\text{DTR-depth}}$	Settling depth	Layer depth at prediction stabilization
$I_{\text{MTP-cos}}$	MTP: cosine distance	Distance between prediction-head representations
$I_{\text{DTR-deep}}$	Deep-settling indicator	Binary depth readout
$I_{\text{EigenScore}}$	EigenScore	Representation dispersion
$I_{\text{Mahalanobis}}$	Mahalanobis distance	Distance in representation space
$B_{\text{MC-TU}}$	MC dropout: TU	Entropy of the mean sampled distribution
$B_{\text{MC-AU}}$	MC dropout: AU	Mean entropy across sampled distributions
$B_{\text{TokUR-TU}}$	TokUR: TU	Total-uncertainty readout
$B_{\text{TokUR-AU}}$	TokUR: AU	Aleatoric-uncertainty readout
$B_{\text{MoE-TU}}$	MoE perturbation: TU	Total uncertainty under routing perturbations
$B_{\text{MoE-AU}}$	MoE perturbation: AU	Conditional entropy under routing perturbations
$B_{\text{LogTokU-Sum}}$	LogTokU: concentration	Sum of Dirichlet concentration parameters
$B_{\text{LogTokU-EU}}$	LogTokU: EU	Evidence-based epistemic-uncertainty readout
$B_{\text{MC-Top1}}$	MC dropout: top-1 probability	Confidence readout from repeated forwards
$B_{\text{TokUR-Top1}}$	TokUR: top-1 probability	Associated confidence readout
$B_{\text{MoE-Top1}}$	MoE perturbation: top-1 probability	Associated confidence readout
$B_{\text{MC-EU}}$	MC dropout: EU	Total entropy minus expected conditional entropy
$B_{\text{TokUR-EU}}$	TokUR: EU	Epistemic-uncertainty readout
$B_{\text{MoE-EU}}$	MoE perturbation: EU	Disagreement under routing perturbations
$B_{\text{LogTokU-AU}}$	LogTokU: AU	Evidence-based aleatoric-uncertainty readout
G_{SelfEval}	Self-evaluated confidence	Prompted confidence judgment

D Mechanism and Estimator Details

This appendix provides the aggregation protocols, mathematical derivations, and additional comparisons supporting Section 4.

D.1 Probability-rank analysis

We analyze probability-rank buckets, Rényi orders, logit profiles, and learned estimators on the 7,216-position UncertainBench. For bucket analysis, we compute correlations within each model and domain family, then aggregate their absolute magnitudes across these groups. The domain families are AO3, arXiv, BBC, bioRxiv, GitHub, and Wikipedia. The bucket analysis measures within-family associations, whereas the baseline comparison computes correlations across domains within each model. For order-scan and held-out-text results, we compute correlations within each model and aggregate their magnitudes across models. All purity values are ratios of the unrounded aggregate correlations. The main bucket plot retains the probability-rank buckets shared by all analyzed models.

D.2 Logit predictions and observed profiles

For a fixed oracle, expected log loss has output-logit gradient

$$\frac{\partial\, \mathbb{E}_{t\sim\bar{P}}[-\log P(t \mid c)]}{\partial z_u} = P(u \mid c) - \bar{P}(u \mid c).$$

This gives a suppressive update direction outside the oracle support and a reinforcing direction for underpredicted reasonable tokens. Parameter coupling prevents interpreting these output-space directions as guaranteed changes in every logit after an actual network update. The proposed proportionality between knowledge-reinforcement frequency and logit separation is a mechanistic hypothesis; the frozen-model contrasts do not measure reinforcement frequency or estimate that scaling relation. We define high- and low-EU groups using the upper and lower thirds of each model's own EU proxy labels and compare their mean logits. The analysis uses effective pre-softmax logits.

D.3 RÉNYI LIMITS AND ORDER SCANS

For a finite strictly positive distribution, write $p_t = P(t \mid c)$ and $V = |\mathcal{V}|$. Rényi entropy is

$$H_\alpha(P) := \frac{1}{1-\alpha} \log \sum_t p_t^\alpha. \tag{13}$$

Its limits give $H_\infty = -\log p_{(1)}$, $H_2 = -\log \sum_t p_t^2$, and $H_1 = -\sum_t p_t \log p_t$. The normalized power weights $w_\alpha(t) = p_t^\alpha / \sum_u p_u^\alpha$ concentrate on tokens with maximum probability as $\alpha \to +\infty$, tend to $1/V$ as $\alpha \to 0$, and concentrate on tokens with minimum probability as $\alpha \to -\infty$. Indeed, $\frac{\mathrm{d}}{\mathrm{d}\alpha} \log \sum_t p_t^\alpha = \sum_t w_\alpha(t) \log p_t$, connecting these weights to the contribution of token log probabilities. Writing $\bar{\ell} = V^{-1} \sum_t \log p_t$, a near-zero expansion gives

$$H_\alpha = \log V + \alpha(\log V + \bar{\ell}) + O(\alpha^2).$$

Hence the first-order limit is

$$\lim_{\alpha \to 0} \frac{\log V - H_\alpha(P)}{\alpha} = -\log V - \frac{1}{V} \sum_t \log p_t = D_{\mathrm{KL}}(U \| P), \tag{14}$$

where U is uniform over the vocabulary; this is the token-level Self-Certainty score (Kang et al., 2025), whereas $H_0 = \log V$ itself is constant. These relationships concern score transformations and rank correlations, not equality of the untransformed scores. The empirical order comparison is reported in Figure 5c; the first-order relation above motivates its connection to Self-Certainty.

D.4 LEARNING AND HELD-OUT EVALUATION

Shared predictor. The model in Section 4.2 combines 57 distribution features with an 8-dimensional model embedding. Its shared backbone maps $65 \to 64 \to 1$, using GELU and dropout 0.1 after the hidden layer. The features comprise eight head log-probability and gap features, six quantiles over the next 120 ranks, ten tail-bucket means, and 33 probability summaries, including cumulative mass and entropy-related statistics. The embedding supplies model identity while the hidden and output transformations share parameters.

The reported optimization uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, 60 epochs, and seed 0. It minimizes mean squared error against reasonable-token mass $1 - \widehat{\mathrm{EU}}$; the uncertainty output is one minus predicted mass. This is equivalent to ordinary squared-error EU supervision under that output transformation. AU proxy labels are not inputs or regression targets. The shared parameters allow pooling training examples across LLMs; a model-specific embedding does not by itself define an estimator for an unseen LLM.

Ridge probe. For feature vector x of bucket mean log probabilities, ridge regression minimizes

$$\frac{1}{N} \sum_{j=1}^{N} (y_j - b - w^\top \widetilde{x}_j)^2 + \lambda \|w\|_2^2,$$

where $\widetilde{x}$ is standardized using training statistics and the intercept is unpenalized. The coefficient analysis uses $|w_i| / \sum_j |w_j|$ within a fit before aggregating across fits. Standardization makes these coefficients comparable in scale, but correlated inputs prevent a causal interpretation of their magnitudes.

Matched held-out comparison. The text-level evaluation reserves half of the source texts for testing and divides the remaining half into five folds. Four folds train each predictor and one serves for validation, giving the $4:1:5$ ratio in the main text. Predictions on the fixed test set are averaged across the five runs. Positions from one source text remain together, preventing adjacent positions from being distributed across training and test partitions. Table 9 reports the matched held-out results. The reference is $|\rho_{EU}| = 0.44758$ on these held-out records, giving $100(0.51580/0.44758 - 1) = 15.2\%$ for the reported improvement. The 0.401 Self-Certainty value comes from the broad baseline evaluation; the improvement above uses the baseline on the same held-out texts as the learned estimators.

Estimator	$\lvert\rho_{EU}\rvert$	$\lvert\rho_{AU}\rvert$	π_{EU} (%)
$\mathcal{M}_{\text{free}}$	0.4476	0.4110	52.13
$\mathcal{M}_{\text{ridge}}$	0.4843	0.3050	61.36
$\mathcal{M}_{\text{train}}$	0.5158	0.2494	67.41

Table 9: Matched held-out results. Correlation magnitudes are aggregated before the purity ratio is computed.

Training domains	Held-out domains	$\mathcal{M}_{\text{free}}$	$\mathcal{M}_{\text{ridge}}$	$\mathcal{M}_{\text{train}}$
3	12	0.4063	0.4417	0.4501
5	10	0.4072	0.4464	0.4562
7	8	0.4147	0.4621	0.4755
10	5	0.4019	0.4576	0.4735
12	3	0.3957	0.4584	0.4784

Table 10: Domain-transfer correlations. Each entry aggregates results across models and 30 partitions, after averaging oriented correlations across held-out domains.

Held-out domains and application scope. The transfer study uses 15 domains, with 30 random domain partitions for each training-domain count $k \in \{3, 5, 7, 10, 12\}$. Training and validation use only the selected training domains; the remaining domains are evaluated without refitting. Within each model and partition, we average the signed, training-oriented Spearman correlations over held-out domains, then aggregate across model–partition records. At $k = 7$, each record includes eight held-out domains. This differs from taking the absolute value separately in every held-out domain. The measured experiment in Figure 6c holds out domains within the benchmark; the proposed application workflow extends the trained predictor to new corpora for the same LLMs. Independent external-corpus and unseen-LLM evaluations remain separate tests.

E Downstream Evaluation Details

The following protocols specify the evaluation units, score aggregation, and complete results for the five applications in Section 5.

E.1 Response and model evaluation

Models and sampling. The response experiments cover 31 benchmarks, with ten responses sampled per question at temperature 1. Table 11 lists the LLMs used for knowledge reliability, response quality, and model capability; Tables 12–14 report the benchmark results. The per-benchmark model coverage depends on response and label availability. Each comparison uses responses with aligned uncertainty scores and correctness labels. Scores are collected on assistant-generated tokens, excluding input prompts and tool outputs. The shared predictor's features, architecture, and optimizer are specified in Appendix D.4; downstream scoring uses the resulting token-level outputs without target-benchmark correctness supervision.

Response aggregation. For a response $w_{1:T}$, let c_j be the prefix before w_j and u_j a token-level EU estimate. Its response-level EU score is $T^{-1}\sum_j u_j$. Shannon entropy is also averaged across positions, while the other probability baselines use

$$U_{\text{PPL}} = \exp\left(-\frac{1}{T}\sum_{j=1}^{T}\log P(w_j \mid c_j)\right), \tag{15}$$

$$U_{\text{MaxProb}} = -\frac{1}{T}\sum_{j=1}^{T}\max_{t\in\mathcal{V}} P(t \mid c_j), \tag{16}$$

$$U_{\text{MaxSurprisal}} = \max_{1\le j\le T}[-\log P(w_j \mid c_j)]. \tag{17}$$

Model	Model
DeepSeek-V4-Flash	Qwen3-8B
GLM-4.5-Base	Qwen3-8B-Base
GLM-4.7	Qwen3.5-0.8B
Kimi-K2.5	Qwen3.5-0.8B-Base
Qwen3-0.6B	Qwen3.5-122B-A10B
Qwen3-0.6B-Base	Qwen3.5-27B
Qwen3-1.7B	Qwen3.5-2B
Qwen3-1.7B-Base	Qwen3.5-2B-Base
Qwen3-14B	Qwen3.5-35B-A3B
Qwen3-14B-Base	Qwen3.5-35B-A3B-Base
Qwen3-235B-A22B	Qwen3.5-4B
Qwen3-30B-A3B	Qwen3.5-4B-Base
Qwen3-30B-A3B-Base	Qwen3.5-9B
Qwen3-32B	Qwen3.5-9B-Base
Qwen3-4B	Qwen3.8-27B
Qwen3-4B-Base	Step-3.5-Flash-Base

Table 11: LLMs used in the knowledge reliability, response quality, and model capability experiments.

The sign for MaxProb makes larger scores indicate greater uncertainty. MaxSurprisal depends on the generated token, not on the least probable token in the vocabulary. PPL and mean surprisal give identical response rankings, but averaging them over responses can produce different model rankings because exponentiation and averaging do not commute.

Reliable knowledge versus lucky success. For model m and question q, write $a_{mq} = k^{-1}\sum_{r=1}^{k} y_{mqr}$ for the observed $\mathrm{avg@k}$, where y_{mqr} is binary correctness. For a pair of models answering the same question, we retain comparisons only if their a_{mq} values differ and each supplies at least one correct response. Every pair of correct responses contributes one comparison. It receives credit one when the model with lower response EU has higher a_{mq}, zero when it has lower a_{mq}, and 0.5 when EU is tied. Within a benchmark, eligible response pairs are weighted equally; the final 78.6% result then gives benchmarks equal weight. The ten sampled responses construct the empirical reliability target; making a comparison requires only one correct response from each model.

Response correctness. Correct and incorrect responses are compared within the same model and question, excluding groups without outcome variation. We report AUROC strength $\max(A, 1-A)$ for AUROC A, a measure of discriminative association whose finite-sample chance expectation can exceed 0.5.

Model capability. Within each benchmark, response-level scores are averaged for each model and correlated with its benchmark accuracy across models. The reported aggregate is

$$\frac{1}{31}\sum_{b=1}^{31}\left|\mathrm{Spearman}_m(\bar{U}_{m,b}, a_{m,b})\right|, \tag{18}$$

where $\bar{U}_{m,b}$ and $a_{m,b}$ denote model-mean uncertainty and accuracy. Tables 12–14 contain all 31 benchmarks for all six displayed estimators, in a common order. Their means reproduce Figure 7 using equal benchmark weights and benchmark-level scores reported to three decimals.

Table 12: Knowledge reliability: accuracy for selecting the model with higher avg@10 from correct responses. Bold entries identify the largest reported value within a benchmark; the final row weights all 31 benchmarks equally.

Benchmark	$\mathcal{M}_{\text{train}}$	$\mathcal{M}_{\text{free}}$	Shannon	PPL	MaxProb	Max Surprisal
agieval	**0.777**	0.745	0.763	0.755	0.746	0.590
aime2024	**0.743**	0.592	0.610	0.609	0.591	0.616
aime2025	**0.818**	0.563	0.616	0.614	0.600	0.570
alpacaeval2	**0.741**	0.574	0.616	0.614	0.609	0.578
arc_challenge	**0.806**	0.721	0.673	0.669	0.662	0.584
arc_easy	0.809	0.787	**0.813**	0.807	0.800	0.687
bbh	**0.791**	0.750	0.615	0.607	0.574	0.571
bigcodebench_full	**0.804**	0.753	0.620	0.574	0.539	0.465
bigcodebench_hard	**0.725**	0.720	0.606	0.587	0.554	0.533
ceval	0.800	0.799	**0.816**	0.798	0.779	0.558
cmmlu	0.792	0.796	**0.816**	0.797	0.778	0.551
commonsenseqa	0.798	0.769	**0.804**	0.798	0.793	0.659
drop	0.753	0.756	**0.780**	0.769	0.764	0.644
global_mmlu_lite	**0.826**	0.793	0.817	0.806	0.787	0.591
gpqa_diamond	**0.763**	0.696	0.732	0.729	0.720	0.577
hellaswag	**0.799**	0.727	0.779	0.773	0.765	0.644
livecodebench	0.702	0.727	**0.754**	0.751	0.747	0.718
longbench_v2	**0.715**	0.693	0.674	0.604	0.651	0.513
math500	**0.884**	0.866	0.796	0.772	0.752	0.542
mbpp	0.868	**0.960**	0.913	0.902	0.871	0.780
minerva_math	**0.879**	0.859	0.787	0.755	0.739	0.484
mmlu	**0.812**	0.754	0.727	0.726	0.713	0.646
mmlu_pro	**0.789**	0.772	**0.789**	0.776	0.768	0.619
nq	0.746	0.711	**0.747**	0.727	0.726	0.581
olympiadbench	**0.853**	0.798	0.752	0.743	0.719	0.615
openbookqa	0.798	0.764	**0.802**	0.796	0.790	0.667
piqa	**0.796**	0.746	0.786	0.781	0.776	0.663
simpleqa	**0.694**	0.669	0.624	0.600	0.585	0.454
triviaqa	**0.762**	0.725	0.755	0.739	0.732	0.603
truthfulqa_mc	**0.756**	0.703	0.753	0.738	0.728	0.536
winogrande	**0.772**	0.720	0.765	0.759	0.753	0.627
Mean	**0.786**	**0.742**	**0.739**	**0.725**	**0.713**	**0.596**

Table 13: Response quality: AUROC strength within the same model and question. Bold entries identify the largest reported value within a benchmark; the final row weights all 31 benchmarks equally.

Benchmark	$\mathcal{M}_{\text{train}}$	$\mathcal{M}_{\text{free}}$	Shannon	PPL	MaxProb	Max Surprisal
agieval	0.572	**0.576**	0.554	0.565	0.548	0.554
aime2024	0.715	**0.748**	0.675	0.668	0.672	0.643
aime2025	0.732	**0.770**	0.678	0.668	0.687	0.697
alpacaeval2	0.565	**0.567**	0.550	0.560	0.550	0.544
arc_challenge	0.554	**0.589**	0.573	0.571	0.562	0.564
arc_easy	0.590	0.601	**0.604**	0.597	0.599	0.560
bbh	0.580	0.587	0.604	0.599	**0.609**	0.594
bigcodebench_full	0.588	**0.627**	0.609	0.616	0.603	0.593
bigcodebench_hard	0.594	**0.627**	0.622	0.622	0.620	0.592
ceval	**0.587**	0.564	0.577	0.565	0.569	0.576
cmmlu	0.581	0.564	0.568	0.565	0.568	**0.590**
commonsenseqa	0.578	**0.587**	0.582	0.586	0.578	0.567
drop	**0.582**	0.559	0.561	0.565	0.576	0.550
global_mmlu_lite	**0.575**	0.561	0.558	0.560	0.559	0.567
gpqa_diamond	0.541	**0.542**	0.532	0.532	0.530	0.522
hellaswag	0.593	0.596	0.585	**0.600**	0.588	0.572
livecodebench	0.678	**0.722**	0.676	0.694	0.683	0.663
longbench_v2	0.537	0.536	0.534	0.534	**0.538**	0.524
math500	0.598	**0.644**	0.621	0.623	0.620	0.637
mbpp	0.689	0.674	0.652	**0.697**	0.621	**0.697**
minerva_math	0.614	**0.672**	0.637	0.623	0.630	0.669
mmlu	0.572	0.566	0.565	0.566	0.558	**0.573**
mmlu_pro	0.538	**0.561**	0.556	0.550	0.557	0.536
nq	0.547	**0.567**	0.559	0.562	0.559	0.531
olympiadbench	0.639	**0.672**	0.632	0.631	0.627	0.593
openbookqa	**0.626**	0.623	0.619	0.621	0.615	0.581
piqa	0.576	**0.580**	0.572	0.569	0.570	0.562
simpleqa	0.551	0.560	**0.578**	0.573	**0.578**	0.570
triviaqa	0.572	**0.584**	0.576	0.579	0.574	0.563
truthfulqa_mc	**0.576**	0.562	0.567	0.571	0.564	0.568
winogrande	**0.581**	0.568	0.570	0.565	0.572	0.562
Mean	**0.594**	**0.605**	**0.592**	**0.593**	**0.590**	**0.584**

Table 14: Model capability: absolute Spearman correlation between model-mean uncertainty and benchmark accuracy. Bold entries identify the largest reported value within a benchmark; the final row weights all 31 benchmarks equally.

Benchmark	$\mathcal{M}_{\text{train}}$	$\mathcal{M}_{\text{free}}$	Shannon	PPL	MaxProb	Max Surprisal
agieval	0.847	0.709	0.831	0.809	**0.850**	0.311
aime2024	0.848	0.800	0.877	0.846	**0.899**	0.431
aime2025	**0.850**	0.735	0.812	0.811	0.824	0.287
alpacaeval2	**0.903**	0.285	0.576	0.721	0.503	0.588
arc_challenge	**0.811**	0.606	0.695	0.736	0.625	0.210
arc_easy	**0.826**	0.625	0.754	0.752	0.751	0.665
bbh	**0.809**	0.702	0.698	0.786	0.553	0.408
bigcodebench_full	**0.821**	0.721	0.574	0.562	0.518	0.447
bigcodebench_hard	**0.825**	0.679	0.800	0.556	0.751	0.603
ceval	**0.850**	0.720	0.839	0.833	0.849	0.412
cmmlu	**0.852**	0.723	0.823	0.804	0.828	0.445
commonsenseqa	**0.825**	0.607	0.757	0.758	0.746	0.549
drop	0.712	0.655	**0.723**	0.710	0.707	0.534
global_mmlu_lite	0.866	0.743	**0.876**	0.864	0.872	0.499
gpqa_diamond	**0.883**	0.676	0.826	0.801	0.839	0.211
hellaswag	**0.753**	0.500	0.700	0.706	0.721	0.563
livecodebench	**0.736**	0.400	0.391	0.391	0.345	0.700
longbench_v2	0.503	0.132	0.471	0.400	0.488	**0.523**
math500	0.684	0.872	0.931	**0.934**	0.923	0.672
mbpp	**0.349**	0.248	0.259	0.243	0.274	0.283
minerva_math	0.776	0.865	0.900	**0.907**	0.894	0.572
mmlu	**0.820**	0.633	0.756	0.739	0.759	0.350
mmlu_pro	**0.911**	0.718	0.858	0.857	0.880	0.404
nq	**0.801**	0.558	0.745	0.698	0.754	0.386
olympiadbench	**0.924**	0.836	0.889	0.877	0.891	0.532
openbookqa	**0.852**	0.611	0.772	0.774	0.781	0.632
piqa	**0.827**	0.574	0.757	0.758	0.763	0.618
simpleqa	**0.746**	0.432	0.667	0.609	0.695	0.200
triviaqa	0.692	0.506	0.675	0.646	**0.695**	0.483
truthfulqa_mc	**0.884**	0.589	0.799	0.792	0.822	0.361
winogrande	**0.813**	0.544	0.703	0.723	0.692	0.475
Mean	**0.794**	**0.613**	**0.733**	**0.723**	**0.726**	**0.463**

E.2 Critical Branching Positions

Branching degree. Let $S(c) \subseteq \mathcal{V}$ be the support of the actual next-token sampler, with $P(t \mid c)$ normalized over that support and zero elsewhere. For each token $t \in \mathcal{V}$, let $Y(t, c)$ denote the final-answer correctness probability under continued sampling after $c \oplus t$. Then $B(c)^2 = \sum_{t \in \mathcal{V}} P(t \mid c)(Y(t, c) - \bar{Y}(c))^2$, where $\bar{Y}(c) = \sum_{t \in \mathcal{V}} P(t \mid c) Y(t, c)$. The degree depends on the sampler, including its temperature and any top-k/top-p truncation. It lies in $[0, 1/2]$ and is zero when the sampler has only one possible next token. When $\bar{Y}(c) \in (0, 1)$, $B(c)/\sqrt{\bar{Y}(c)(1 - \bar{Y}(c))}$ provides an additional normalized measure.

Search and labeling. The evaluation uses 189 SimpleQA Verified (Haas et al., 2025) responses from Qwen3.6-35B-A3B, comprising 253,370 candidate positions, 913 labeled positives, and 252,457 positions treated as negatives. For a sampled response prefix $w_{1:t}$, the labeling procedure estimates $q_t = \Pr(\text{correct final answer} \mid w_{1:t})$ by repeated suffix generation and compares q_t with q_{t-1}. Progressive binary search first probes sparse prefix landmarks with 100 suffix samples, recursively bisects intervals with large correctness changes, and verifies surviving candidate positions with 200 and then 500 samples per prefix. Borderline cases receive up to 2,000 samples; the final thresholds are $|q_t - q_{t-1}| \geq 0.03$ and a standardized change of at least 2. The 913-position evaluation set contains 497 correctness increases and 416 decreases.

Sampling cost. Search, verification, and counterfactual sampling for the 189 evaluated responses use 5,632,500 suffix rollouts, or $5{,}632{,}500/913 \approx 6{,}169$ per labeled position (approximately 6,200 in the main text). This cost is scoped to the evaluated responses and excludes initial response screening and earlier runs.

This realized-path change is a proxy for branching, rather than a direct estimate of $B(c)$. For exact correctness probabilities and a token drawn from the same sampler, $\mathbb{E}[(q_t - q_{t-1})^2 \mid c_t] = B(c_t)^2$. A single realized branch can nevertheless miss large variation among alternative tokens. The branch-wise extension below measures that variation directly; the reported AUROCs use the path-based labels.

Branch-wise estimation algorithm. For direct estimation of the branching degree, a branch-wise procedure explicitly evaluates alternative continuations:

1. Recover the sampler's candidate support at each position. Remove singleton supports, whose branching degree is exactly zero.
2. Use coarse prefix-correctness probes and subsequent local refinement to prioritize a candidate set. This is a search heuristic, not exhaustive certification.
3. For each candidate position, force every supported token and sample suffixes to obtain pilot estimates of $Y(t, c)$. Allocate a separate confirmation budget to the surviving positions and branches.
4. Estimate branching variance from fresh confirmation samples, with the finite-sample noise correction below, and quantify uncertainty by resampling branch outcomes.
5. Rank positions by estimated branching degree; retain branch probabilities, outcome estimates, sample counts, and uncertainty intervals for verification.

For independent confirmation estimates $\hat{Y}(t, c)$ from fixed counts $m(t, c) > 1$ for each $t \in S(c)$, an unbiased estimator of $B(c)^2$ is

$$\begin{aligned}\widehat{B^2}(c) = &\sum_{t \in S(c)} P(t \mid c)(\hat{Y}(t, c) - \hat{\bar{Y}}(c))^2 \\ &- \sum_{t \in S(c)} P(t \mid c)(1 - P(t \mid c)) \frac{\hat{Y}(t, c)(1 - \hat{Y}(t, c))}{m(t, c) - 1}, \\ \hat{\bar{Y}}(c) = &\sum_{t \in S(c)} P(t \mid c)\hat{Y}(t, c).\end{aligned} \tag{19}$$

We retain signed variance estimates for diagnostics and use $\sqrt{\max(0, \widehat{B^2}(c))}$ for display.

Ranking evaluation and categories. AUROC is computed within each response and then averaged over responses. Each evaluated response contains at least one detected positive, with unmarked positions treated as negatives. The evaluation therefore measures agreement with detected path changes; the search may leave other branching positions unmarked. The aligned comparison excludes methods requiring an auxiliary multi-token-prediction head. It includes full-vocabulary probability scores, MC dropout, and layerwise disagreement scores. Category-wise evaluation treats positions of one category as positives and all remaining positions in that response as negatives; only responses containing that category contribute.

Estimator	Mean AUROC	SD	Pooled AUROC
$\mathcal{M}_{\text{train}}$	0.6825	0.1526	0.6800
$\mathcal{M}_{\text{free}}$	0.6777	0.1525	0.6747
Layer-JSD	0.6476	0.1497	0.6579
MC dropout	0.6339	0.1477	0.6300
Shannon	0.6255	0.1495	0.6282
MaxProb	0.6172	0.1438	0.6194
PPL	0.6151	0.1436	0.6181

Table 15: Branching-position detection on the same 189 responses. The mean weights responses equally; SD is variation across responses, not a confidence interval. Pooled AUROC ranks all retained positions together.

Category	Count	$\mathcal{M}_{\text{train}}$	$\mathcal{M}_{\text{free}}$	Shannon	Layer-JSD	MC	PPL
Factual knowledge	372	0.703	0.693	0.598	0.623	0.604	0.584
Elaboration path	331	0.655	0.647	0.648	0.667	0.650	0.644
Sentence boundary	124	0.708	0.720	0.660	0.685	0.647	0.654
Response formatting	60	0.570	0.570	0.572	0.599	0.605	0.562
Reasoning pivot	11	0.414	0.461	0.296	0.384	0.411	0.287
Other surface form	15	0.701	0.651	0.600	0.626	0.593	0.583

Table 16: Category-wise mean AUROC. Each category is evaluated on responses containing at least one positive of that category, with remaining positions as negatives. MC denotes MC dropout; Layer-JSD uses the negative mean divergence between intermediate-layer and final-layer predictions on the selected vocabulary support.

Factual knowledge includes entity-name recall, numerical-value selection, and characterization choices. Other categories are elaboration path, sentence boundary, response formatting, reasoning pivot, and surface form. The reasoning-pivot and surface-form groups contain only 11 and 15 positives, respectively, limiting the interpretation of their category-level results. The paired response-level AUROC difference between $\mathcal{M}_{\text{train}}$ and full-vocabulary Shannon entropy is 0.0570, with a bootstrap 95% interval of $[0.0408, 0.0729]$. The difference between $\mathcal{M}_{\text{train}}$ and $\mathcal{M}_{\text{free}}$ is smaller (0.0048, interval $[-0.0014, 0.0111]$); these results do not establish a significant difference between the two estimators.

Token-level PPL and MaxProb. At a generated token w_t with prefix c_t, token-level PPL is $1/P(w_t \mid c_t)$; its logarithm $-\log P(w_t \mid c_t)$ gives the same ranking for AUROC. Sampled-token surprisal yields a mean AUROC of 0.6151. MaxProb instead ranks positions by $-\max_{v \in \mathcal{V}} P(v \mid c_t)$, equivalently $-\log \max_{v \in \mathcal{V}} P(v \mid c_t)$, and yields 0.6172 on the same 189 responses.

E.3 Pre-trained model potential

Data and target. The experiment uses continued-pretraining checkpoints from internal development, with different data mixtures, hyperparameters, and training-data volumes. All checkpoints undergo the same post-training procedure. Their potential labels are post-training performance on SWE-bench Verified (Jimenez et al., 2024), SWE-bench Pro (Deng et al., 2025), and SWE-bench Multilingual (Yang et al., 2025). The pre-trained checkpoints are evaluated on the same 7,597 trajectories generated by Claude Opus 4.6 (Anthropic, 2026). Uncertainty is aggregated over generated tokens, excluding prompts and tool outputs. Within each benchmark, Spearman correlation is computed across checkpoints and the benchmark correlations are then averaged. The bucket analysis preserves the sign; the method comparison reports mean correlation magnitudes.

Rank localization. Table 17 specifies the bucket intervals and signed correlations. These intervals differ from the doubling buckets in Section 4.1: the first three are ranks 1, 2–4, and 5–16. The top-1 probability is positively associated with potential, whereas mean log probabilities in the subsequent buckets are negatively associated. The strongest bucket-level association occurs at ranks 5–16 ($\rho = -0.939$).

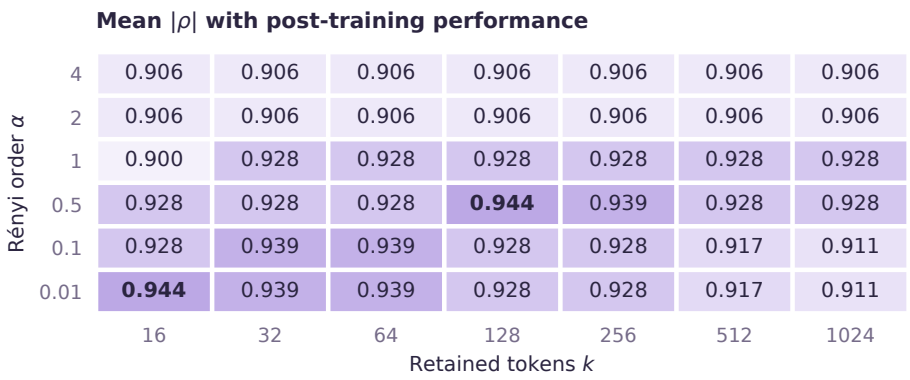


Figure 14: Mean absolute Spearman correlation for top-k Rényi entropy. Bold values mark both best grid settings at the recorded three-decimal precision.

Top-k Rényi entropy. Let $T_k(c)$ contain the k most probable tokens and $Z_k(c) = \sum_{t \in T_k(c)} P(t \mid c)$. Define $P^{(k)}(t \mid c) = P(t \mid c)/Z_k(c)$ on $T_k(c)$ and zero outside it. The score is

$$H_\alpha(P^{(k)}) = \frac{\log \sum_{t \in T_k(c)} P(t \mid c)^\alpha - \alpha \log Z_k(c)}{1 - \alpha}. \tag{20}$$

Top-k selection restricts the score to the head, while $0 < \alpha < 1$ reduces the dominance of its highest-probability token. The resulting token scores are averaged over the same trajectory positions as the other estimators.

Figure 14 shows the full grid of six orders and seven support sizes. Both $(\alpha, k) = (0.5, 128)$ and $(0.01, 16)$ attain a mean absolute correlation of 0.944 at three-decimal precision; the main text uses the former, whose signed correlation is -0.944. The grid was examined on the same checkpoints used for evaluation, so the comparison is exploratory rather than a held-out hyperparameter validation. Table 18 reports the baseline values used in the main text. HE-SNR (Wang et al., 2026) relates target-token probability to head-distribution entropy on selected high-entropy positions.

Bucket	Rank interval	Mean ρ
1	1	$+0.906$
2	2–4	-0.917
3	5–16	-0.939
4	17–64	-0.933
5	65–128	-0.928
6	129–256	-0.917
7	257–512	-0.911
8	513–1K	-0.917
9	1K–2K	-0.889
10	2K–4K	-0.822
11	4K–8K	-0.776
12	8K–16K	-0.742
13	16K–33K	-0.710
14	33K+	-0.728

Table 17: Signed correlations of bucket mean log probability with post-training performance. K abbreviates approximately one thousand ranks; the final interval extends to the end of the vocabulary.

Estimator	Mean $\lvert\rho\rvert$
Head Rényi ($\alpha = 0.5$, $k = 128$)	0.944
Shannon	0.922
HE-SNR	0.909
PPL	0.909
$\mathcal{M}_{\text{train}}$	0.834
$\mathcal{M}_{\text{free}}$	0.741

Table 18: Mean absolute Spearman correlation with post-training performance across the three SWE-bench benchmarks.

E.4 READING THE DOWNSTREAM RESULTS

The tasks use distinct units of analysis: pairs of correct responses for knowledge reliability, responses within a model–question group for correctness, models within a benchmark for capability, positions within a response for branching, and checkpoints within a benchmark for potential. Their reported values therefore answer different questions and should not be compared as interchangeable measures of EU accuracy. The branching labels measure detected changes along sampled paths, while the potential experiment explores correlations on a small internal checkpoint set. Together with the benchmark-level result tables, these distinctions specify the evidence behind the strengths and limitations discussed in Section 5.

F LIMITATIONS AND FUTURE DIRECTIONS

Signal sources and learning within LLMs. Although our baseline comparison includes methods based on model internals, our mechanistic analysis and estimator development focus primarily on next-token predictive distributions. The extent to which hidden representations, attention patterns, or intermediate computations provide complementary EU signals remains open. Our learned estimator reads signals from existing LLMs without updating their parameters. Training LLMs themselves to represent and expose informative EU signals is a further direction beyond the present study. Pretraining or post-training objectives could encourage this behavior through auxiliary prediction heads or explicit supervision that distinguishes knowledge gaps from the multiplicity of reasonable continuations. Whether these approaches improve estimation accuracy, EU purity, and transfer across models and domains requires further investigation. UNCERTAINBENCH provides a common testbed for evaluating such advances against token-level EU and AU proxy labels.

Broader downstream applications. The five tasks studied here cover response assessment, model evaluation, and reasoning analysis, but represent only part of the possible uses of token-level EU. Further applications include training-data selection and curriculum design, detection of erroneous or corrupted text, prioritization of examples for annotation, and decisions about retrieval, tool use, or additional inference compute. These settings pose different requirements for aggregating uncertainty and acting on it. For example, a correct but unfamiliar passage may elicit high EU, so data selection and error detection require distinguishing learning opportunities from evidence of incorrect content. Our response-quality and potential-prediction results show that stronger token-level EU estimation does not automatically yield a better downstream criterion. Future studies should therefore assess both the quality of the uncertainty signal and the effects of the resulting decisions on task performance, data efficiency, and computational cost.

G Related Work

Research on uncertainty in language models spans probability calibration, uncertainty decomposition, hallucination detection, and selective generation. These directions share an interest in reliable predictions, but differ in the quantity being estimated and the evidence used to evaluate it (Baan et al., 2023; Geng et al., 2024). We organize their relationship to this work around four questions: what confidence means, how prediction granularity matters, how uncertainty definitions relate, and what existing evaluation protocols establish.

G.1 Uncertainty and confidence in LLMs

Confidence requires a specified event. In calibration research, confidence is a forecast of the probability that a prediction is correct; calibration asks whether predictions assigned a given confidence are correct at the corresponding frequency (Guo et al., 2017). For generative language models, the event may instead be producing a particular token, giving a correct answer, or making a supported factual claim. These events have different probabilities. Jiang et al. (2021) study the calibration of answer probabilities in question answering and show that strong task performance does not ensure calibrated confidence. Consequently, a large $P(t \mid c)$ expresses the model's preference for token t, and becomes evidence about correctness only in relation to a specified task and evaluation protocol.

Confidence can be elicited through different channels. Kadavath et al. (2022) examine self-evaluation of answer correctness and whether a model can predict that it knows an answer. Lin et al. (2022) train models to express confidence in words, while Tian et al. (2023) find that verbalized confidence can be better calibrated than conditional answer probabilities for models trained with human feedback. These findings support treating verbal reports, predictive probabilities, and internal representations as distinct sources of confidence signals. They also motivate the possibility, discussed in Appendix F, of training LLMs to expose more informative uncertainty signals themselves. Successful calibration of an answer-level report, however, does not by itself establish that the report separates EU from AU at individual generation positions.

Predictive concentration and knowledge are different properties. A concentrated predictive distribution can favor an unreasonable continuation, while a knowledgeable model can distribute probability across several reasonable alternatives. This distinction is central to uncertainty decomposition (Hüllermeier & Waegeman, 2021; Hou et al., 2024). Under our proxy labels, $1 - \widehat{\mathrm{EU}}(P, c) = \sum_{t \in R(c)} P(t \mid c)$ is the probability of a reasonable next token according to the annotated set. It therefore gives a precise local reliability interpretation without requiring one reasonable token to dominate. This quantity differs from both maximum token probability and the probability of a correct final answer. Our correlation-based evaluation measures association with this local target; probability calibration and downstream correctness remain additional evaluation questions.

G.2 Token-level and response-level uncertainty

Aggregation has a probabilistic basis and a task-dependent role. Autoregressive factorization expresses a response's log probability as a sum of conditional token log probabilities. The entropy chain rule likewise relates sequence entropy to conditional token entropies after averaging over possible prefixes. Malinin & Gales (2021) develop token- and sequence-level uncertainty measures within this framework, and Zhang et al. (2025) formulate token-to-response aggregation with expectation-level relationships. These identities clarify how specified probabilistic quantities connect across levels. They do not establish that an estimator with better response-correctness discrimination must more accurately recover EU at each position. In particular, a score averaged along one generated response depends on its realized prefixes, and averaging discards information about where uncertainty occurs.

Meaning and token choice introduce different variation. Semantically equivalent answers can use different words and therefore have different token probabilities. Kuhn et al. (2023) address this issue by grouping generated answers according to semantic equivalence and estimating entropy over meanings. Farquhar et al. (2024) develop this approach for detecting confabulations, a class of incorrect generations that vary across samples, and explicitly distinguish this target from systematic

errors. Semantic aggregation removes variation among paraphrases, but it does not necessarily remove uncertainty due to genuinely different reasonable answers. Indeed, Farquhar et al. (2024) explicitly leave the separation of aleatoric and epistemic uncertainty outside their analysis. Semantic uncertainty and our token-level EU proxy labels thus address related reliability questions at different levels of representation.

Not every token contributes equally to the downstream decision. Duan et al. (2024) show that weighting tokens and sentences by semantic relevance improves response-level uncertainty estimation. Their results motivate studying aggregation separately from the quality of the underlying local signal. Our applications make this distinction explicit: the preferred estimator can change between recovering token-level EU and discriminating correct responses (Section 5.1). Similarly, branching degree measures how next-token choices change eventual correctness, which also depends on subsequent reasoning (Section 5.2). A useful response or branching score may therefore require both local uncertainty and information about the token's role in the larger computation.

G.3 Relating uncertainty definitions through a common reference

Predictive, aleatoric, and epistemic uncertainty. The distinction between uncertainty inherent in the prediction problem and uncertainty attributable to the learner is well established (Kendall & Gal, 2017; Hüllermeier & Waegeman, 2021). Its interpretation depends on the information available and the output being predicted. For language generation, AU can reflect multiple valid continuations or an underspecified question; Hou et al. (2024) use input clarification to study the latter. Calling AU irreducible means irreducible for the specified conditioning information: supplying a clarification changes the prediction problem and can reduce its ambiguity. We adopt this conditional view throughout, with the next token as the output and c as the available context.

EU as excess predictive risk. Lahlou et al. (2023) define EU through the excess risk above an ideal predictor, including mismatch caused by model misspecification. Under log loss, this yields the reference-distribution formulation used in Section 2.1: AU is $H(\bar{P})$ and EU is $D_{\mathrm{KL}}(\bar{P}\|P)$, while predictive TU is $H(P)$. For a fixed oracle, the quantity that equals AU plus EU is the expected log loss under $\bar{P}$, as derived in equation 4. It is not generally $H(P)$. Keeping these quantities distinct allows a confidently wrong model to have low predictive entropy and substantial oracle-relative EU.

Bayesian disagreement and evidential measures. Bayesian methods describe uncertainty over possible predictive models; MC dropout and deep ensembles provide practical approaches to obtaining predictive variation (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). Within a Bayesian joint model, expected conditional entropy and mutual information sum to posterior predictive entropy. As shown in Appendix A.1, this is also the posterior expectation of our oracle-based AU and EU when the evaluated prediction is the posterior predictive mean. The connection requires these probabilistic assumptions: agreement among a finite set of approximate models need not exclude a shared error relative to a fixed oracle (Lahlou et al., 2023). Evidential approaches offer another representation; for example, Ma et al. (2025) derive a Dirichlet distribution from logits and use it to quantify uncertainty and response reliability. Such scores are interpretable within their constructions, while their correspondence to the common oracle-based target remains an empirical question.

A shared target for comparing different estimators. This perspective unifies the comparison by specifying the context, prediction space, reference distribution, and loss, rather than assuming that every quantity called uncertainty is numerically equivalent. UncertainBench operationalizes that comparison through reasonable continuation sets. Its $\widehat{\mathrm{AU}} = \log|R(c)|$ is the entropy of a uniform pooled reference, and $\widehat{\mathrm{EU}}$ is a transformed lower-bound proxy retaining probability mass outside the reasonable set. The latter does not measure all within-set discrepancies captured by the full oracle KL (Appendix A.3). An important antecedent is Xiao & Wang (2021), who define next-step hallucination probability as the total probability of contextually false tokens. Our EU proxy shares this probability-mass interpretation; the oracle-risk derivation, pooled reasonable-set construction, and accompanying AU proxy labels make it an operational basis for the comparison studied here. These explicit labels provide a common basis for asking whether probability-based,

internal, Bayesian, or elicited scores track local unreasonable-continuation risk and how strongly they also respond to the multiplicity of reasonable continuations.

G.4 How token-level uncertainty estimators have been evaluated

Response-level quality and selection are widely used targets. A substantial strand of work evaluates uncertainty through the quality of completed generations. LM-Polygraph compares methods using prediction rejection: outputs are ranked by uncertainty, and the quality of retained outputs is assessed with ROUGE-L or BERTScore (Fadeeva et al., 2023). SAR aggregates relevance-weighted signals and evaluates discrimination of response correctness (Duan et al., 2024). LogTokU evaluates response reliability on TruthfulQA, including AUROC against correctness labels defined using BLEURT (Ma et al., 2025). Self-Certainty is evaluated through best-of-N response selection (Kang et al., 2025), while TokUR assesses incorrect-response detection and selection in mathematical reasoning and other generation tasks (Zhang et al., 2025). These protocols establish useful downstream behavior, but their evaluation targets are responses even when the estimator computes a score at every token.

Direct and fine-grained evaluation also has important precedents. Malinin & Gales (2021) evaluate token-error detection in speech recognition using reference alignments, while explicitly noting why token-error labeling is ill-posed for translation with valid alternative wordings. Xiao & Wang (2021) connect predictive uncertainty to hallucination in image captioning and data-to-text generation. For free-form LLM text, Ahdritz et al. (2024) use a larger model as a proxy for the unknown ground-truth distribution and train probes to identify positions where it would be confident. This is a particularly relevant precedent for distinguishing token-level epistemic from aleatoric-like uncertainty; its labels depend on the larger model's predictive entropy. Other work evaluates finer units of factuality: Lookback Lens detects contextually unsupported spans (Chuang et al., 2024a), and FActScore measures support for atomic factual claims (Min et al., 2023). These studies demonstrate that fine-grained evaluation is feasible, while targeting different notions of error or uncertainty.

The distinction is between evaluation targets, not only their granularity. A token or span error label concerns the realized continuation; a distribution-level label concerns the alternatives available before that continuation is chosen. One observed reference token does not specify all reasonable next tokens, and one correct response does not reveal the risk assigned to unreasonable alternatives at each prefix. Likewise, a larger model's entropy and an annotation of factual support provide different information from separately constructed EU and AU proxy labels. UncertainBench pools and validates reasonable alternatives and evaluates probability mass over that set for each model, making these distinctions accessible to a common comparison. Its contribution is therefore a shared operational testbed for token-level EU and AU proxies, complementing existing calibration, error-detection, and selective-generation evaluations.

Implications for interpreting estimator performance. We report association with EU alongside association with AU because a score can perform well at predicting output quality while also responding to benign variation in valid continuations. EU purity summarizes this relative association; it is not a variance decomposition or a proof of statistical independence. Conversely, success on these local labels does not guarantee superior downstream decisions, which depend on aggregation, the correctness criterion, and the role of individual tokens. This motivates the progression from direct measurement and estimator evaluation to mechanistic analysis and separate application studies in the main text.