BEYOND BINARY CORRECTNESS: CONFIDENCE ESTIMATION FOR CONTINUOUSLY GRADED RESPONSES IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Confidence estimation predicts the correctness of generated outputs of LLMs. Most existing approaches typically use an arbitrary threshold over a continuous quality metric, such as syntactic or semantic similarity against ground truth, to determine whether an LLM response is correct. In this paper, we propose modeling confidence as a probability distribution over *continuous* correctness scores, thereby remaining agnostic to threshold choice, avoiding information loss from binarization, and enabling rich uncertainty estimates such as confidence variance and credible intervals for robust decision making. We propose Beta regression and piecewise uniform regression approaches as distributional methods, showing theoretically that Beta regression achieves superior sample efficiency and faster convergence for smooth distributions with at most one interior mode. We also conduct an empirical study on several datasets across diverse tasks to analyze and compare methods with each other, as well as traditional point-based confidence estimation baselines, demonstrating their relative suitability and advantages.

1 INTRODUCTION

When large language models (LLMs) generate responses to complex queries, how confident should we be in their outputs? This question is critical for deploying LLMs in domains where incorrect outputs can have severe consequences. *Confidence estimation* for LLMs has emerged as a critical research area, with methods developed to predict the *correctness* of generated outputs via information from model logits, self-consistency, verbalized uncertainty, or supervised learning from features. Tasks such as question answering, summarization, translation, and open-ended generation naturally evaluate output quality against ground truth using continuous metrics, including syntactic similarity (ROUGE, BLEU), semantic similarity, domain-specific rubrics, or even correctness scores from a frontier LLM. Geng et al. (2024) survey the state-of-the-art.

Existing confidence estimation methods treat correctness as binary ($y \in \{0, 1\}$), with confidence typically interpreted as the probability of a response being correct ($P(y = 1)$). When quality metrics are naturally continuous, this requires practitioners to define "correct" by thresholding quality at an arbitrary cutoff point τ. This approach suffers from several limitations. First, **thresholding discards information**: responses scoring 0.51 and 0.95 are treated identically as "correct" when $\tau = 0.5$, losing all nuance about response quality. Second, binary models are **threshold-dependent**: choosing $\tau \in \{0.3, 0.5, 0.7\}$ yields different models with different performance characteristics, and changing the threshold requires retraining. Third, binary confidence provides only $P(y \geq \tau)$, offering **limited uncertainty quantification** compared to a full probability distribution over correctness scores. Finally, this framework is unable to support **uncertainty-aware decision-making** – in practice, leveraging uncertainty in predictions (e.g., intervals [0.70, 0.75] vs [0.40, 0.95]) may be beneficial for determining which responses require review. Figure 1 illustrates some of these limitations. Panel (A) shows how different threshold choices fundamentally alter the prediction task, with no principled basis for selecting among them. Panel (B) demonstrates how binary methods discard distributional information: two responses may yield similar binary probabilities, yet have different underlying uncertainty that binary classification is unable to capture.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

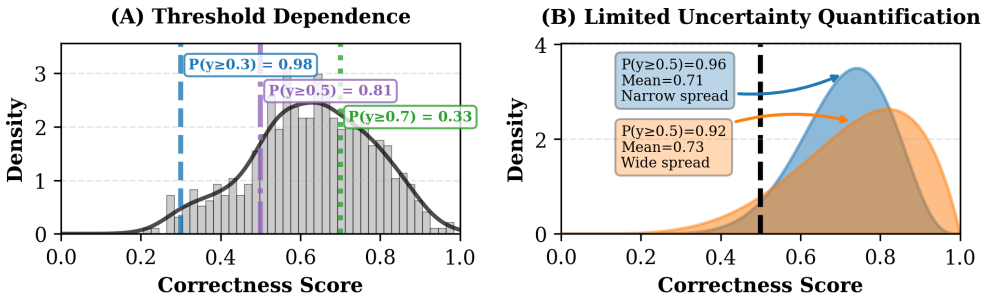


Figure 1: Limitations of binary confidence estimation via thresholding. **(A)** Different threshold choices yield different binary classification tasks with varying proportions classified as "correct". There is no principled way to select the threshold, yet it fundamentally determines the prediction task. **(B)** Binary methods discard distributional information. Two responses with similar binary probabilities (0.96 vs. 0.92) can have different uncertainty (narrow vs wide distributions).

In this paper, we propose modeling confidence as a probability distribution over *continuous* correctness scores ($y \in [0, 1]$) using Beta regression and its variant as our primary approaches. We additionally develop two complementary distributional methods – multi-threshold binary regression and quantile regression – to serve as principled baselines. This distributional framework preserves the full granularity of quality metrics, remains agnostic to threshold choice (enabling post-hoc computation of $P(y \geq \tau)$ for any τ), provides rich uncertainty estimates including mean, variance, and credible intervals, and naturally supports nuanced decision policies that account for uncertainty-aware correctness. Furthermore, they require limited training data and are therefore practical and easy to adopt in real-world applications. We operate in the setting where there is black-box (or at best, gray-box) access to the LLM whose responses are being assessed, i.e. we assume no access to model weights but potential availability of token logit information. Our primary contributions are as follows:

- **Distributional confidence estimation for continuous correctness.** We extend confidence estimation from binary classification to distributional regression for continuous correctness – a problem setting that has not been previously studied: specifically, we propose two variants of Beta regression, multi-threshold binary regression, and quantile regression for predicting full distributions.
- **Theoretical characterization of method trade-offs.** We provide the first formal analysis guiding method selection for distributional confidence estimation: Beta regression is more sample efficient and converges faster for smooth distributions with at most one interior mode, while non-parametric methods are preferable for internally multimodal or irregular distributions.
- **Comprehensive empirical validation.** Experiments on diverse tasks and datasets illustrate when and why Beta (or an extension) achieves better calibration and distributional fit than other models.

Together, these contributions provide a unified framework that spans parametric and non-parametric distributional estimation, with theoretical guidance validated empirically across datasets and tasks.

2 Related Work

Confidence estimation has been explored since before the advent of LLMs, such as for question-answering systems (Kamath et al., 2020), but has gained renewed attention. Prior work distinguishes between *uncertainty* estimation – quantifying model variability in responses, which is a function of the input – and *confidence* estimation – predicting correctness of a specific response, which is a function of both input and output (Lin et al., 2024b). These are closely related but separate endeavors in *uncertainty quantification* (UQ) for LLMs, surveyed by several recent articles (Fadeeva et al., 2023; Geng et al., 2024; Xia et al., 2025; Liu et al., 2025). Confidence estimation methods for LLMs fall into two broad categories: *white-box* and *black-box*. White-box approaches assume access to the LLM's internal components, such as model weights, logits or embeddings, while black-box methods infer confidence solely from the model's outputs. Methods that use token logits and no other internal components are often referred to as *gray-box* – a subset of white-box methods.

Common white-box approaches use token-level probabilities as a signal for model confidence (Malinin & Gales, 2021; Duan et al., 2023; Lin et al., 2024a), consistent with broader UQ findings that logits are informative of reliability (Kuleshov et al., 2018; Ott et al., 2018; Desai & Durrett, 2020; Vazhentsev et al., 2023). Kadavath et al. (2022)elicit verbalized True/False confidence via token probability; others use internal embeddings (Ren et al., 2023) or attention layers (Kadavath et al., 2022; Burns et al., 2023; Li et al., 2023; Azaria & Mitchell, 2023; Vazhentsev et al., 2025). Black-box methods typically measure consistency across multiple generations via natural language inference (Kuhn et al., 2023) or syntactic/semantic similarity (Lin et al., 2024b; Farquhar et al., 2024; Gao et al., 2024; Bhattacharjya et al., 2025) or use verbalized self-evaluation (Mielke et al., 2022; Xiong et al., 2024; Lin et al., 2022). Recent work has explored combining white- and black-box UQ (Chen & Mueller, 2024; Shrivastava et al., 2023), as well as fitting prediction models using correctness training data (Kapoor et al., 2024; Li et al., 2024; Liu et al., 2024; Bhattacharjya et al., 2025; Yaldiz et al., 2025; Vazhentsev et al., 2025). Importantly, most methods treat confidence as representing a binary notion of correctness, i.e., measuring whether a response is correct or not.

Our approach differs fundamentally from post-hoc calibration of binary classifiers, e.g., isotonic regression (Zadrozny & Elkan, 2002) and temperature scaling (Guo et al., 2017), which adjusts probabilities for a fixed binary outcome rather than performing distributional regression. We also differ from quality estimation (Specia et al., 2018; Rei et al., 2020), which regresses continuous quality but yields only point estimates, and from conformal prediction (Angelopoulos & Bates, 2021), which gives distribution-free coverage guarantees without modeling $p(y|x)$. Our distributional approach enables richer queries: variance, arbitrary quantiles, threshold-agnostic probabilities $P(y \geq \tau)$ without retraining, and interpretable credible intervals that reflect learned uncertainty structure.

3 Distributional Confidence Estimation

We formulate confidence estimation for LLMs as predicting $p(y|x)$ where $y \in [0, 1]$ is continuous correctness and x represents the extracted features. Unlike prior work that treats this problem as binary classification, distributional approaches preserve full granularity and enable threshold-agnostic uncertainty quantification. $p(y|x)$ should be interpreted as the conditional distribution of correctness over the population of inputs sharing similar feature values x, capturing aleatoric variability not explained by x (e.g., grading sensitivity, phrasing effects), rather than a distribution over repeated generations for a single input. This is the same statistical object modeled by any conditional regression with predicted variance, such as heteroscedastic Gaussian or quantile regression; our contribution is applying this lens to LLM correctness. We propose two Beta regression-based models and two piecewise uniform models. We first describe the feature extraction process common to all methods.

3.1 Feature Extraction

We compile features (denoted x) from four complementary categories established in prior work; full feature definitions and complete per-category lists with descriptions are provided in Appendix A.

Query Features capture task difficulty via query length, question type, and structural markers, motivated by evidence that question type and length correlate with difficulty (Rajpurkar et al., 2018; Longpre et al., 2021). **Response Features** capture linguistic uncertainty via response length, word/sentence counts, and hedge-word markers (e.g., "maybe," "possibly"), following work showing these verbalized cues and response length correlate with model uncertainty (Vincze et al., 2008; Lin et al., 2022; Kamath et al., 2020). **Token Probability Features** use mean token log-probability from generation, a standard confidence signal where lower values indicate higher uncertainty (Guo et al., 2017; Xiao & Wang, 2021; Kadavath et al., 2022; Malinin & Gales, 2021). **Similarity Features** exploit the consistency hypothesis, which posits that similar outputs across repeated generations indicate higher confidence (Xiao et al., 2025), using pairwise similarities across multiple samples per query, used individually or aggregated by mean, max, or standard deviation (Wang et al., 2023; Manakul et al., 2023; Lin et al., 2024b; Bhattacharjya et al., 2025).

These features are established signals from prior work rather than our contribution, which is around combining all four categories for regression over continuous correctness. Features are standardized to zero mean and unit variance using training-set statistics; see Appendix A for full details.

3.2 BETA MODELS

Beta regression models $y|x \sim \text{Beta}(\alpha(x), \beta(x))$ where shape parameters are a function of features x. We parameterize via mean $\mu(x) = \text{logit}^{-1}(x^\top \mathbf{w})$ and precision ϕ, with $\alpha = \mu\phi$, $\beta = (1-\mu)\phi$. Parameters are estimated via maximum likelihood, and a well-known transformation is used to ensure support on $(0, 1)$ (Smithson & Verkuilen, 2006). While Beta regression can model heteroscedastic precision $\phi(x)$, we use constant precision for stability: variable precision exhibited some numerical instability in cross-validation. We propose the following two variants of Beta models.

Beta Regression. Standard Beta regression assumes $y \sim \text{Beta}(\alpha(x), \beta(x))$ with constant precision ϕ. This 2-parameter model is sample-efficient and suitable for smooth distributions with at most one interior mode. We fit via maximum likelihood using L-BFGS-B optimization.

Hurdle-Beta. We empirically observed that some benchmarks, such as for short-form question answering, have a concentration at $y = 1$ (e.g., from exact matches in BERTScore). For such datasets, we consider a modification of the one-inflated mixture model (Ospina & Ferrari, 2012) which we refer to as the *Hurdle-Beta* model, motivated by hurdle models from count data (Mullahy, 1986):

$$p(y|x) = \begin{cases} \pi_1(x) & \text{if } y \in [\tau_h, 1] \\ (1 - \pi_1(x)) \cdot \text{Beta}(y|\alpha(x), \beta(x)) & \text{if } y \in (0, \tau_h) \end{cases}$$

where τ_h is a boundary-adjacent hurdle threshold to account for numerical precision and near-boundary concentration, and $\pi_1(x)$ is the boundary probability predicted by a calibrated logistic classifier. Note that τ_h is a preprocessing parameter analogous to bandwidth selection in kernel density estimation; it handles numerical artifacts near $y = 1$ but does not define task-level correctness, so the framework remains threshold-agnostic and users can compute $P(y \geq \tau)$ for any τ without retraining, regardless of the τ_h used during model fitting. We use $\tau_h = 0.99$ in our experiments.

This two-component mixture explicitly models the point mass at perfect correctness while maintaining Beta's flexibility for intermediate scores.

3.3 PIECEWISE UNIFORM MODELS

Piecewise uniform models construct the conditional distribution $y|x$ as a histogram with uniform density within bins. Unlike parametric Beta models, these non-parametric approaches make no assumptions about distributional shape, at the cost of increased sample complexity. We propose two variants that differ in how they define bin edges.

Multi-Threshold Binary Regression, Multi-threshold binary regression extends prior work that uses logistic regression to estimate $P(y \geq \tau|x)$ for a single threshold τ, primarily using similarity-based features for x (Bhattacharjya et al., 2025). We generalize this by training k binary classifiers at fixed thresholds $\tau_1 < \tau_2 < ... < \tau_k$ to predict $P(y \geq \tau_i|x)$. Each classifier uses logistic regression with binary cross-entropy loss on labels $\mathbb{1}(y \geq \tau_i)$. The piecewise uniform distribution is defined by:

$$p_i(x) = P(\tau_{i-1} \leq y < \tau_i|x) = P(y \geq \tau_{i-1}|x) - P(y \geq \tau_i|x)$$

yielding bin probabilities, where $\tau_0 = 0$ and $\tau_{k+1} = 1$. By combining predictions across thresholds, we obtain a full distributional estimate with $k + 1$ bins at fixed edges, enforcing the monotonicity constraint via isotonic regression (Zadrozny & Elkan, 2002).

Quantile Regression. Quantile regression estimates the conditional quantile function $Q_\tau(x) = \inf\{y : P(Y \leq y|x) \geq \tau\}$ for multiple quantile levels. Each quantile is learned independently by minimizing the pinball loss:

$$L_\tau(y, \hat{q}) = \begin{cases} \tau(y - \hat{q}) & \text{if } y \geq \hat{q} \\ (1 - \tau)(\hat{q} - y) & \text{if } y < \hat{q} \end{cases}$$

The CDF is constructed via linear interpolation between predicted quantiles $\hat{q}_{\tau_1}(x) < ... < \hat{q}_{\tau_n}(x)$, which induces a piecewise uniform distribution with sample-adaptive bin edges. Unlike multi-threshold regression, bin locations vary across samples to match local distributional structure. Another key difference is that quantile regression models the inverse CDF (quantile function) instead of the CDF directly; it derives the CDF via interpolation.

3.4 Theoretical Comparison of Distributional Methods

We compare the approximation quality and sample complexity of Beta regression and multi-threshold piecewise uniform regression, addressing when Beta's parametric structure is preferable to the non-parametric piecewise alternative. Let $p(y|x)$ denote the true conditional distribution, and let $\mathcal{F}_{\text{Beta}}$ and $\mathcal{F}_{\text{PU}(k)}$ denote the families of Beta distributions and k-bin piecewise uniform distributions, respectively. For a distribution class $\mathcal{F}$, define approximation error $\delta_{\mathcal{F}}(p) := \inf_{q \in \mathcal{F}} D_{KL}(p\|q)$. Results rely on subsets of the following assumptions:

- **(A1) Smoothness:** Continuous first and second derivatives.
- **(A2) Unimodality:** At most one local maximum in $(0, 1)$.
- **(A3) Bounded density:** $\exists\, c_{\min}, c_{\max} > 0$ such that $c_{\min} \leq p(y|x) \leq c_{\max}$ for all $y \in [\epsilon, 1-\epsilon]$, some $\epsilon > 0$.

Theorem 1 (Beta vs. Piecewise Uniform Trade-offs)**.** *Consider regression with n i.i.d. samples.*

Part A (Approximation): *Under (A1), piecewise uniform approximation error decays as $O(k^{-4})$ and achieves universal approximation as $k \to \infty$; under multimodality ($m \geq 2$ interior modes), Beta regression incurs irreducible approximation error $\delta_{\mathcal{F}_{Beta}}(p) \geq C_m > 0$.*

Part B (Finite-Sample Estimation): *Under (A3), Beta regression's total error decomposes as $\delta_{\mathcal{F}_{Beta}}(p) + O(1/n)$; under (A1) and (A3), piecewise uniform's total error is bounded by $O(k^{-2}) + O(k \log k/n)$.*

Part C (Regime Comparison): *Combining Parts A–B: Beta regression achieves lower total error for smooth, unimodal, bounded distributions when $\delta_{\mathcal{F}_{Beta}}(p) \ll C\|p''\|_\infty^2/k^4$, $k \leq 7$, and $n \in [100, 10000]$; piecewise uniform is preferable for multimodal distributions ($m \geq 2$) when $n \geq Ck^3$; and asymptotically, as $k, n \to \infty$ with $k^2 \ll n$, piecewise uniform dominates whenever $\delta_{\mathcal{F}_{Beta}}(p) > 0$.*

Theorem 1 formalizes a practical trade-off: Beta regression is more sample-efficient for smooth, unimodal distributions, while piecewise methods offer universal approximation for irregular or multimodal ones. This extends to conditional regression $p(y|x)$: for d-dimensional features, Beta learns $2d$ parameters with variance $O(d/n)$, while multi-threshold learns kd parameters with variance $O(kd/n)$, preserving Beta's efficiency advantage when $n \gtrsim kd$. Full proofs, exact constants, and the conditional-regression extension are in Appendix B. We next show empirically that LLM correctness distributions often favor Beta regression.

4 Experimental Setup

Datasets. We consider 7 datasets across 4 task categories:

- **Short-form question answering** (HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017)): Extractive and multi-hop QA datasets requiring 2-5 word factoid answers.
- **Summarization** (CNN Dailymail (Nallapati et al., 2016), XSum (Narayan et al., 2018)): News article summarization requiring compression into 50-100 word summaries.
- **Long-form generation** (Biographies (Min et al., 2023)): Multi-paragraph biographical narratives about notable figures, requiring sustained coherence and factual accuracy.
- **Code generation** (BigCodeBench (Zhuo et al., 2025)): Python function synthesis evaluated by functional correctness on comprehensive unit test suites.

We subsample 1000 queries each for all question answering and summarization datasets. Biographies and BigCodeBench have 157 and 1130 queries, respectively.

Models. For QA, summarization, and long-form generation, we generate responses using open-source models Granite4.0 32B and LLaMA3.3 70B, which represent typical LLMs in their model classes. For BigCodeBench, we use Ministral 8B and Qwen2.5-Coder 7B.

Correctness Scores. In this work, correctness for text generation tasks is primarily gauged by semantic similarity with ground truth using BERTScore (Zhang et al., 2020), which is standard in recent UQ literature. Unlike lexical metrics (e.g., exact match, ROUGE), BERTScore captures

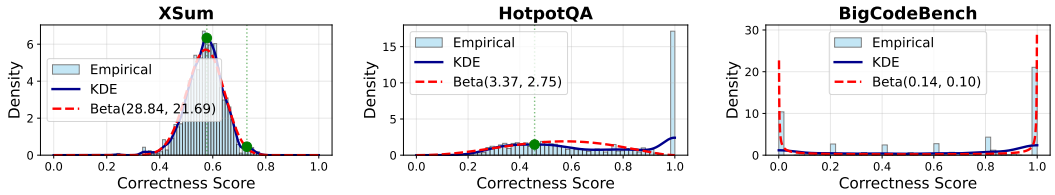


Figure 2: Marginal correctness distributions at zero temperature for XSum, HotpotQA, and BigCodeBench (left to right), spanning unimodal, boundary-concentrated, and multimodal shapes. In the middle panel, Beta is fit only below 0.99, matching Hurdle-Beta.

semantic equivalence through contextual embeddings, recognizing paraphrases as correct while producing continuous [0,1] scores suitable for distributional modeling, and its correlation with human correctness judgments is empirically validated in Zhang et al. (2020). We conduct an ablation study with ROUGE-L and F1 later in an experiment (Section 5.2), confirming our conclusions are consistent across correctness metrics. Code generation is evaluated by unit test pass rates (Chen et al., 2021).

Correctness distributions vary substantially across tasks, as illustrated in Figure 2: XSum is unimodal (well-suited to parametric Beta modeling); HotpotQA shows boundary concentration ($\sim$28% of responses $\geq$ 0.99), the scenario Hurdle-Beta targets; and BigCodeBench is multimodal, since correctness is near-discrete (e.g., many outputs pass exactly 2 of 5 unit tests, yielding a mode at 0.4). We use these three datasets in our detailed results (Table 1) to span this range of distributional shapes. Figure 2 illustrates marginal shapes for motivation; our actual evidence for conditional fit is provided by the favorable results in Section 5.

Evaluation Metrics. We consider a suite of metrics that provide complementary views of model quality, including calibration and discriminative ability of estimates of expected correctness, as well as measures of distributional fit and uncertainty quantification.

- **Mean Absolute Error (MAE):** Measures point prediction accuracy as the average absolute difference between predicted means and true correctness scores. We also report Selective-MAE, which restricts MAE to the most reliable predictions as ranked by lowest predicted uncertainty, to assess the practical value of predictive uncertainty for triage and abstention.
- **Continuous Expected Calibration Error (ECE):** Uses the continuous variant of ECE (Guo et al., 2017) that bins predicted means (rather than class probabilities) and compares with empirical means. We use 10 bins for experiments.
- **Interval Score (IS):** Evaluates predicted credible interval quality at a specified level (e.g., 90%) by penalizing both interval width (sharpness) and coverage violations (Winkler, 1972; Gneiting & Raftery, 2007).
- **Continuous Ranked Probability Score (CRPS):** A proper scoring rule measuring the distance between predicted cumulative distribution and empirical distribution (Gneiting & Raftery, 2007).

We use CRPS as a metric for assessing distributional fit instead of the more popular negative log-likelihood (NLL) (Murphy, 1973), because we need to compare a mixture distribution (Hurdle-Beta) with fully continuous distributions; this requires comparison over CDF and not PDF. Detailed definitions of these metrics (as well as additional ones) are provided in Appendix C.

Point Estimate Baselines. We compare our distributional approaches against point estimate baselines that predict only the mean correctness without modeling the full distribution.

- **Linear Regression:** Standard least-squares regression.
- **Random Forest:** Ensemble of 100 trees where intervals are derived from the empirical distribution across component tree predictions (Mentch & Hooker, 2016).
- **Calibrated Data-Free:** Data-free confidence scores, such as from token logits (Malinin & Gales, 2021), aggregated similarity (Lin et al., 2024a), or semantic cluster sizes (Kuhn et al., 2023); these are calibrated via isotonic regression (Zadrozny & Elkan, 2002).

Calibrated data-free methods are only considered in Appendices to limit the number of point baselines. Despite reasonable performance on some point metrics, we exclude them from the main comparison

Table 1: Confidence estimation results for 4 evaluation metrics across 3 datasets and 2 LLMs with generations at temperature zero using stochastic decoding. LLM #1 and #2 are Llama-3.3 70B and Granite-4.0 32 B for XSum and HotpotQA, and Ministral 8B and Qwen2.5-Coder 7B for BigCodeBench. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta when no concentrated mass appears near perfect correctness, like in XSum. MT refers to multi-threshold.

		LLM #1				LLM #2			
	Method	**MAE↓**	**ECE↓**	**IS@90%↓**	**CRPS↓**	**MAE↓**	**ECE↓**	**IS@90%↓**	**CRPS↓**
XSum	Beta Reg	**0.053**±.002	**0.004**±.003	**0.296**±.037	**0.038**±.002	**0.054**±.003	0.006±.003	**0.316**±.030	**0.039**±.002
	Hurdle Beta	-	-	-	-	-	-	-	-
	MT Binary Reg	0.056±.004	0.013±.006	0.489±.011	0.042±.002	0.057±.003	0.014±.006	0.495±.011	0.043±.002
	Quantile Reg	**0.053±.002**	0.005±.003	0.318±.036	0.039±.002	**0.054**±.002	0.006±.002	0.338±.038	0.041±.002
	Linear Reg	**0.053**±.002	**0.004**±.004	N/A	N/A	0.055±.003	**0.005**±.003	N/A	N/A
	Random Forest	0.055±.002	0.015±.004	0.368±.031	N/A	0.057±.003	0.012±.005	0.355±.039	N/A
Hotpot QA	Beta Reg	0.219±.010	0.085±.012	0.824±.027	0.154±.006	0.223±.010	0.095±.016	0.834±.025	0.157±.008
	Hurdle Beta	**0.199**±.007	**0.029**±.012	**0.762**±.023	**0.127**±.004	**0.200**±.008	**0.027**±.014	**0.772**±.006	**0.129**±.006
	MT Binary Reg	0.206±.008	0.032±.014	0.779±.032	0.133±.003	0.205±.006	0.029±.002	0.781±.017	0.132±.004
	Quantile Reg	0.213±.010	0.040±.015	0.766±.031	0.136±.005	0.211±.006	0.046±.011	0.780±.039	0.136±.003
	Linear Reg	0.212±.010	0.037±.021	N/A	N/A	0.211±.006	0.038±.012	N/A	N/A
	Random Forest	0.200±.010	0.070±.011	1.018±.064	N/A	0.189±.009	0.063±.022	1.034±.040	N/A
BigCodeBench	Beta Reg	0.360±.013	0.052±.020	**1.000**±.000	0.212±.007	0.375±.017	0.047±.012	**1.000**±.000	**0.221**±.005
	Hurdle Beta	0.355±.015	0.049±.022	1.075±.005	0.213±.009	0.371±.018	0.062±.023	1.063±.011	0.223±.005
	MT Binary Reg	0.350±.014	**0.031**±.012	1.164±.016	0.216±.010	0.365±.015	**0.037**±.012	1.178±.035	0.225±.006
	Quantile Reg	**0.329**±.020	0.144±.023	1.001±.003	**0.211**±.010	**0.354**±.011	0.132±.016	**1.000**±.000	0.222±.007
	Linear Reg	0.345±.015	0.034±.010	N/A	N/A	0.362±.015	0.038±.007	N/A	N/A
	Random Forest	0.351±.011	0.057±.010	1.021±.013	N/A	0.364±.019	0.077±.025	1.011±.031	N/A

since they lack full distributional predictions and cannot be evaluated on metrics like CRPS or NLL; we only compare them on MAE, ECE, and IS (for Random Forest) in Appendix D.2.

5 Experiments

In this section, we share some results, analyses, and ablations from an empirical study. For our experiments, we generate 5 samples each across temperatures $\{0.0, 0.25, 0.5, 0.75, 1.0\}$ using sampling decoding, although not all generations are utilized for all experiments. Unless otherwise mentioned, similarity-based features for regression are obtained using Jaccard, and thresholds for piecewise uniform methods are fixed to $\{0.1, 0.3, 0.5, 0.7, 0.9\}$.

5.1 Main Results

Table 1 presents our main results across three representative datasets spanning generation tasks and distributional characteristics: XSum (summarization, unimodal distribution), HotpotQA (multi-hop QA, boundary concentration at high correctness), and BigCodeBench (code generation, multimodal distribution from discreteness). We report performance for two LLMs per dataset. We evaluate methods on four complementary metrics: MAE (point accuracy), ECE (calibration), Interval Score @90% (interval reliability), and CRPS (distributional quality).

On XSum's unimodal distribution, Beta and Hurdle-Beta perform identically due to automatic fallback, often achieving best CRPS, ECE, and Interval Score. In contrast, HotpotQA's boundary concentration reveals Hurdle-Beta's advantage: it substantially outperforms standard Beta on CRPS and ECE, demonstrating the value of explicit boundary mass modeling. Across both datasets, Multi-threshold shows poorer calibration and Interval Scores than Hurdle Beta, while Quantile is consistently competitive. On BigCodeBench, piecewise uniform methods excel: Multi-threshold achieves best ECE while Quantile attains best CRPS and MAE, demonstrating non-parametric flexibility when Beta's unimodality assumption is violated. However, Quantile's calibration suffers dramatically, indicating overconfident predicted means despite strong point predictions. Across all datasets, we observe that method choice primarily impacts uncertainty quantification rather than mean predictions.

To summarize performance across all datasets, Table 2(a) reports win rates for each method across three complementary key metrics that capture distinct aspects of performance: MAE measures point prediction accuracy, ECE quantifies calibration of probabilistic predictions, and CRPS evaluates

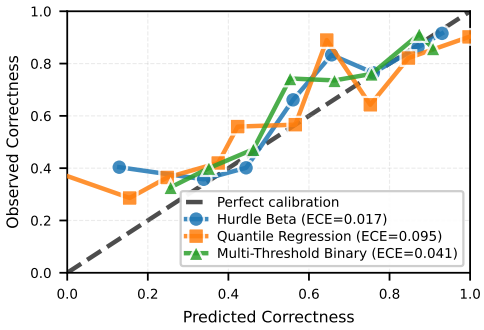
Dataset	H-Beta	MT	Quantile	Linear†	RF†
HotpotQA	2.5/3	0/3	0/3	0/2	0.5/2
NaturalQ	1.5/3	1/3	0/3	0.5/2	0/2
TriviaQA	1.5/3	0/3	1.5/3	0/2	0/2
Biographies	2/3	0/3	0/3	1/2	0/2
CNN/DM	1.5/3	0/3	1/3	0.5/2	0/2
XSum	2/3	0/3	0.5/3	0.5/2	0/2
BigCodeBench	0.5/3	1/3	1.5/3	0/2	0/2
Total Wins	11.5/21	2/21	4.5/21	2.5/14†	0.5/14†
Win Rate	**55%**	10%	21%	18%	4%

(a) Method wins across datasets on MAE, ECE, and CRPS. Points are shared for ties (performance within 1% of top). †Linear and RF baselines do not provide CRPS (1/3 metrics).

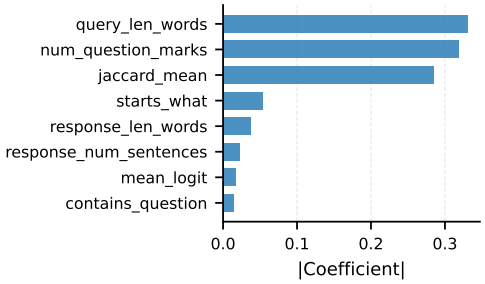
	MAE ↓			CRPS ↓		
Corr.	Beta	Quantile	MT	Beta	Quantile	MT
BERTScore	**0.047**	**0.047**	0.051	**0.037**	0.038	0.044
ROUGE-L	**0.048**	0.049	0.091	**0.036**	0.038	0.065
F1	0.066	0.066	**0.064**	**0.049**	**0.049**	0.051
Temp.	Beta	Quantile	MT	Beta	Quantile	MT
0.25	**0.042**	0.043	0.044	**0.030**	0.032	0.041
0.50	**0.043**	**0.043**	0.045	**0.031**	0.033	0.042
1.0	**0.039**	0.040	0.048	**0.028**	0.031	0.041

(b) Method comparison across correctness metrics and temperature on Biographies, for metrics MAE and CRPS. Generations at temp. 0.0 are used for the ablation on correctness metrics, and BERTScore correctness is used for the ablation on temp.

Table 2: (a) Method wins across datasets over 3 evaluation metrics; (b) Method comparison across correctness metrics and temperature on the Biographies dataset.



(a) Calibration curves on TriviaQA. Points on the curve show binned predictions (x-axis) versus observed correctness (y-axis).



(b) Feature importance for Beta regression on the XSum dataset, ranked by absolute coefficient magnitude. Feature coefficients represent impact on logit-transformed mean correctness.

Figure 3: (a) Calibration curves on TriviaQA; (b) Feature importance for Beta regression on XSum.

overall distributional quality as a proper scoring rule. We do not consider IS as it tends to be highly correlated with CRPS. We report only Hurdle-Beta rather than standard Beta regression, as Hurdle-Beta automatically reduces to Beta when insufficient boundary samples exist, making it a generalization. Wins are shared equally when methods perform within 1% of the top score (e.g., a two-way tie awards 0.5 wins to each method). Hurdle-Beta achieves the highest overall win rate (55%), substantially outperforming alternatives, demonstrating its robustness across diverse datasets. A comprehensive outline of results over all datasets and evaluation metrics is provided in Appendix D.

5.2 Ablations and Analyses

Ablation: Correctness Metric and Temperature. To investigate robustness to correctness metrics and temperature, we repeat experiments on Biographies using Granite-4.0 32B with three metrics – BERTScore (semantic), ROUGE-L (lexical n-gram), and F1 (token-level) – and temperatures $\{0.25, 0.5, 1.0\}$. Table 2(b) shows consistent relative performance across metrics and temperatures. For BERTScore and ROUGE-L, Beta (and Hurdle-Beta) achieves best CRPS, with Quantile competitive and Multi-Threshold (MT) substantially worse; for F1, all methods perform similarly. MT's dramatic degradation on ROUGE-L suggests binary classifiers struggle with fine-grained lexical patterns. This demonstrates that our framework is *metric-agnostic*, modeling $p(y|x)$ for any continuous correctness score. While we use BERTScore for its semantic sensitivity, practitioners can substitute task-appropriate metrics (lexical, rubric-based, or LLM-judged).

Calibration. Figure 3(a) compares calibration curves for all LLama 3.3-70B generations (across temperatures) for TriviaQA. Hurdle-Beta (ECE = 0.017) tracks the diagonal closer than other methods,

Table 3: Selective-MAE@50% for Hurdle-Beta vs. Random Forest.

Method	XSum	HotpotQA	BigCodeBench
Hurdle-Beta	**0.051**	**0.166**	0.333
Random Forest	0.053	0.181	**0.322**

indicating well-calibrated predictions, although it is underconfident near correctness prediction 0.1 and 0.6. Quantile Regression (ECE = 0.095) exhibits more systematic underconfidence, while Multi-Threshold Binary achieves intermediate calibration (ECE = 0.041). This demonstrates that modeling the full Beta can provide superior calibration compared to non-parametric alternatives.

Feature Importance. We illustrate the interpretability of regression models to study feature importance. For simplicity, here we consider the mean similarity with other responses as a similarity-based feature, rather than all pairwise similarities. Figure 3(b) shows the top features from Beta regression on XSum summarization ranked by absolute coefficient magnitude using generations from Llama3.3 70B across all temperatures. Two characteristics of the query, which in this case is the input paragraph to be summarized, dominate: the number of words and the question mark count are the strongest predictors, indicating that question-like and longer inputs yield more predictable correctness, whereas syntactic features show weaker effects. The results also validate prior work on similarity-based confidence estimation since mean similarity is a top predictor. The modest coefficient magnitudes (all $|\beta| < 0.5$) reflect the inherent difficulty of predicting summarization quality.

Selective Prediction. A natural question is whether modeling the full distribution provides value beyond a simpler point-estimate-with-uncertainty baseline. A metric such as Selective-MAE requires a per-sample uncertainty signal besides the point prediction itself; Linear Regression, which outputs only a point estimate, cannot support this evaluation at all, underscoring that even basic selective prediction requires some notion of predictive uncertainty. Among methods that do provide such a signal, we compare Selective-MAE@50% (i.e., MAE on the 50% most reliable predictions, ranked by lowest predicted uncertainty; Appendix C) for Hurdle-Beta against Random Forest's ensemble variance. Table 3 shows Hurdle-Beta achieves lower selective error on XSum and HotpotQA, where correctness is smooth or boundary-concentrated, consistent with Theorem 1; on multimodal BigCodeBench, Random Forest is comparable, matching the theorem's prediction that non-parametric alternatives close the gap under multimodality. The practical benefit of distributional modeling is therefore concentrated where our theory predicts it.

6 Conclusion

In this paper, we have introduced new methods for estimating distributional confidence in continuously graded LLM responses. Unlike previous work that considers confidence estimation as a binary classification problem, we model confidence as a probability distribution over continuous correctness scores and propose Beta regression-based techniques to estimate this distribution. In addition, we develop two complementary methods: multi-threshold binary regression and quantile regression. Our theoretical analysis shows that the proposed Beta regression methods are more sample efficient and converge faster for smooth distributions, while piecewise methods are more suitable for multimodal or irregular distributions. We conduct extensive experiments over a variety of tasks using open-source LLMs, validating that Beta regression methods achieve superior performance with fewer samples.

Our methods naturally inherit **limitations** from the chosen correctness metrics. Parametric Beta assumptions can fail for severely multimodal distributions, though Hurdle-Beta and piecewise methods provide robustness. Practitioners should examine empirical distributions (as in Figure 2) before method selection. For heavily discrete or irregular distributions, piecewise methods or more flexible models may be preferred. Promising future directions include doing a deeper dive around understanding task-dependent beneficial features for distributional regression, exploring more expressive models such as neural networks, studying adaptive methods that select between parametric and non-parametric approaches based on detected distributional characteristics, and pursuing practical real-world applications such as around human-AI collaboration.

AI USE STATEMENT

In this work, we used generative AI tools to help with some methodology related idea generation and for polishing parts of the paper. Additionally, we used generative AI tools to assist with the Python implementation of the proposed methods and plotting figures. We have reviewed all AI-assisted code by running unit tests for smaller and contained functionalities. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The primary risk associated with this work is the potential over-reliance of confidence estimates in high-stakes deployment settings, e.g., medical, legal, or safety-critical applications. It should be understood that a well-calibrated confidence distribution on the tasks and models studied here provides no guarantee under distribution shift, and practitioners should not mistakenly treat predicted correctness distributions as guarantees rather than statistical estimates subject to error outside their validated domain. We emphasize that our proposed methods are intended to support and not replace human oversight, and that calibration should be re-verified on any new target domain before deployment. A related concern is that our extracted features and correctness labels (e.g., BERTScore, derived from contextual embeddings) may reflect biases already present in the underlying LLMs or evaluation models. While our method does not introduce new sources of bias, it also does not correct for pre-existing ones, so confidence estimates inherit any such biases without additional fairness guarantees across subpopulations or query types. This work does not involve human subjects, crowdsourcing, or new data collection, and no IRB approval was required. All experiments use publicly available open-source LLMs and existing benchmark datasets (SQuAD 2.0, CoQA, TriviaQA, HotpotQA, Natural Questions, CNN/DailyMail, XSum, Biographies/FActScore, BigCodeBench) under their existing licenses.

REPRODUCIBILITY STATEMENT

We have taken several steps to support reproducibility of our results. Full model formulations for Beta regression, Hurdle-Beta, multi-threshold binary regression, and quantile regression are given in Section 3, with complete assumptions, theorem statements, and proofs for our theoretical results in Section 3.4 and Appendix B, including the extension to conditional regression. Full feature definitions and counts across all four feature categories are provided in Appendix A. Dataset details, LLMs used, correctness score definitions, and evaluation metric formulas (including Continuous-ECE, Interval Score, CRPS, and Selective-MAE) are described in Section 4 and Appendix C, and extended results across all seven datasets, both LLMs, and all evaluation metrics are reported in Appendix D.2 to allow full verification of the summary results in Section 5. All datasets used are publicly available under their existing licenses, and all LLMs used are open-source and publicly accessible, as noted in our Ethics Statement.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Debarun Bhattacharjya, Balaji Ganesan, Junkyu Lee, Radu Marinescu, Katsiaryna Mirylenka, Michael Glass, and Xiao Shou. SIMBA UQ: Similarity-based aggregation for uncertainty quantification in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15880–15894, 2025.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5186–5200, August 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer, 2001.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*, 2023.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. LM-Polygraph: Uncertainty estimation for language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, 2023.

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

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the European Chapter of the Association for Computational Linguistics EACL)*, pp. 2336–2346, 2024.

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 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, June 2024.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

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

Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous Univariate Distributions*, volume 2. John Wiley & Sons, New York, 2nd edition, 1995.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Doddsand Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *preprint arXiv: 2207.05221 [cs.CL]*, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Volodymyr Kuleshov, Peter Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2796–2804, 2018.

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

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *preprint arXiv: 2306.03341 [cs.CL]*, 2023.

Yukun Li, Sijia Wang, Lifu Huang, and Li-Ping Liu. Graph-based confidence calibration for large language models. *arXiv preprint arXiv:2411.02454*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10351–10368, 2024a.

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

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*, 2024.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6107–6117, 2025.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7052–7063, 2021.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Stefan Riezler and Yoav Goldberg (eds.), *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, August 2016.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Raydonal Ospina and Silvia LP Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, 2012.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Analyzing uncertainty in neural machine translation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3956–3965, 2018.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, 2020.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. Llamas know what GPTs don't show: Surrogate models for confidence estimation. *preprint arXiv: 2311.08877 [cs.CL]*, 2023.

Michael Smithson and Jay Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71, 2006.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162, 2018.

Aad W Van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1430–1454, Toronto, Canada, 2023. Association for Computational Linguistics.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2246–2262, 2025.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 38–45, 2008.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, pp. 1–25, 1982.

Robert L Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. *arXiv preprint arXiv:2503.00172*, 2025.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katsiaryna Mirylenka, Nhan H. Pham, Michael Glass, and Junkyu Lee. The consistency hypothesis in uncertainty quantification for large language models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, number 244, pp. 4636–4651, 2025.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2734–2744, 2021.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. Do not design, learn: A trainable scoring function for uncertainty estimation in generative LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering, 2018. URL `https://arxiv.org/abs/1809.09600`.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, Thong Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, Binyuan Hui, Niklas Muennighoff, David Lo, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. BigCodeBench: Benchmarking code generation with diverse function calls and complex instructions. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

A Features for Regression

We extract features across four categories to capture different aspects of model responses:

Query Features. Characteristics of the input query:

- Query length (characters and words)
- Number of question marks
- Question type indicators (starts with what, how, why, when, where, who)

Response Features. Characteristics of the generated response:

- Response length (characters, words, sentences)
- Uncertainty markers (count of words like "maybe," "possibly," "might," "could," "perhaps," "uncertain," "not sure," "probably," "likely," "unclear")
- Punctuation features (contains question mark, number of exclamation marks)

Token Probability Features. Internal model signal from token logits:

- Mean log probability across generated tokens

Similarity Features. Computed from pairwise comparisons between multiple model-generated samples for each query:

- Pairwise similarities: All pairwise similarities with other samples using a choice of similarity metric (Jaccard, Rouge1, Rouge2, RougeL, sentence-Bert cosine similarity)
- Aggregated similarity: Optionally, use one of the following aggregations of pairwise similarities: mean, variance, or maximum

In total, there are 9 query features, 6 response features, 1 token probability feature, and $m-1$ similarity features (if all pairwise similarities are used) when there are m generated samples in a query. Note that using all individual pairwise similarities as features requires the number of generated samples per query to be constant across all examples to maintain a fixed-size feature vector during training and inference.

B Theorem Proof Outline

We begin by providing some supporting lemmas on density estimation and nonparametric regression that have been established in prior work.

Lemma 1 (MLE Convergence for Beta Regression)**.** *Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. samples with $Y_i|X_i \sim$ Beta$(\alpha(X_i), \beta(X_i))$ where $\alpha(x), \beta(x)$ are parameterized by d coefficients. Under regularity conditions (identifiability, smoothness, bounded Fisher information). If θ^* is the true parameter, then the maximum likelihood estimator $\hat{\theta}_n$ satisfies:*

$$\mathbb{E}[D_{KL}(p_{\theta^*} \| p_{\hat{\theta}_n})] = \frac{d}{2n} + o(1/n)$$

Proof. This follows from standard parametric MLE theory under regularity conditions. See Van der Vaart (1998) Theorem 5.39 for general parametric models, and Wong & Shen (1995) for Beta regression specifically. □

Lemma 2 (Piecewise Constant Approximation Rate)**.** *Let $p : [0,1] \to \mathbb{R}_+$ be a density with $\|p''\|_\infty < \infty$. Partition $[0,1]$ into k equal bins and let q_k be the piecewise constant approximation minimizing $\|p - q_k\|_2^2$. Then, for $C \approx 1/12$:*

$$\|p - q_k\|_2 \leq \frac{C\|p''\|_\infty}{k^2}$$

Proof. This is a result from approximation theory for piecewise constant functions. The proof uses Taylor expansion within each bin of width $h = 1/k$ to bound the local approximation error as $O(h^2\|p''\|_\infty)$, then integrates to obtain the L^2 error. See Yang & Barron (1999) Theorem 1 for the general information-theoretic rate, or Devroye & Lugosi (2001) for the approximation-theoretic perspective. □

Lemma 3 (Piecewise Histogram Sample Complexity)**.** *Let p be a density on $[0,1]$ with $\|p''\|_\infty < \infty$ and $c_{\min} \le p(y) \le c_{\max}$. Let $\hat{q}_k$ be the k-bin histogram estimator from n i.i.d. samples. Then, for C_1, C_2 that depend on $c_{\min}, c_{\max}$:*

$$\mathbb{E}[D_{KL}(p\|\hat{q}_k)] \le \frac{C_1\|p''\|_\infty^2}{k^2} + \frac{C_2 k \log k}{n}$$

Proof. The approximation term $O(1/k^2)$ follows from Lemma 2. The estimation term arises from multinomial sampling: the histogram assigns probability $\hat{p}_j = n_j/n$ to bin j based on counts n_j. Standard concentration inequalities for multinomial distributions yield variance $O(p_j/n)$ per bin. The KL divergence $\sum_j p_j \log(p_j/\hat{p}_j)$ requires control over small probabilities, introducing the $\log k$ regularization factor. See Devroye & Lugosi (2001) Chapter 20, Theorem 2, or Yang & Barron (1999) for information-theoretic bounds. □

Lemma 4 (Beta Unimodality and Multimodal Approximation Failure)**.** *The Beta family has the following properties:*

(i) Every Beta(α, β) with $\alpha, \beta > 1$ has a unique interior mode at $y^* = \frac{\alpha-1}{\alpha+\beta-2}$.

(ii) For any density p on $[0,1]$ with $m \ge 2$ well-separated interior modes, there exists $C_m > 0$ that depends on the mode separation and relative heights, such that:

$$\inf_{q \in \mathcal{F}_{Beta}} D_{KL}(p\|q) \ge C_m$$

Proof. **Part (i):** The Beta density is $p(y;\alpha,\beta) \propto y^{\alpha-1}(1-y)^{\beta-1}$. Taking the derivative:

$$\frac{d}{dy}\log p(y) = \frac{\alpha-1}{y} - \frac{\beta-1}{1-y} = 0 \implies y^* = \frac{\alpha-1}{\alpha+\beta-2}.$$

The second derivative is strictly negative at y^*, confirming it is a unique maximum. For $\alpha \le 1$ or $\beta \le 1$, the mode is at the boundary.

Part (ii): For a density p with interior modes at $y_1 < y_2$, any unimodal Beta distribution q cannot simultaneously match the local structure at both modes. The Beta will fit well near one mode but must underestimate the other, since a unimodal density cannot create the valley between modes without reducing height at the peaks. Let $\epsilon = \min\{|y_1 - y^*|, |y_2 - y^*|\} > 0$ be the separation. By Pinsker's inequality, $D_{\mathrm{KL}}(p\|q) \ge \frac{1}{2}\|p-q\|_1^2$. The L^1 distance between the bimodal p and any unimodal Beta q is bounded below by the mass misplaced near the modes, yielding $\|p-q\|_1 \ge C'_m$ for some constant depending on mode properties. Thus $D_{\mathrm{KL}}(p\|q) \ge C_m = (C'_m)^2/2 > 0$. This argument can be made rigorous using continuity and compactness of the parameter space; see Johnson et al. (1995) Chapter 25 for Beta properties and Devroye & Lugosi (2001) for lower bounds on approximation of multimodal densities. □

Proof Outline of Theorem **??**. The proof follows from the four supporting lemmas.

Part A: Approximation Error Bounds. (i) Piecewise uniform rate: This follows directly from **Lemma 2** (piecewise constant approximation theory). For a density p with bounded second derivative on $[0,1]$, partitioning into k equal-width bins yields approximation error $O(1/k^2)$. The constant $C \approx 1/12$ arises from Taylor expansion of the squared L^2 error within each bin.

(ii) Universal approximation: Immediate from **Lemma 2** with weaker smoothness assumptions. By density of piecewise constant functions in $L^2[0,1]$, any continuous density can be approximated arbitrarily well as $k \to \infty$.

(iii) Beta limitations: This follows from **Lemma 4** (Beta unimodality). The Beta family contains only unimodal densities on $(0,1)$. For any multimodal density p with $m \ge 2$ interior modes, the best

Beta approximation must place its single mode between the true modes, leading to poor fit at both modes and irreducible error $C_m > 0$.

Part B: Finite Sample Estimation Error. **(i) Beta regression:** Under regularity conditions (A3)-(A4), maximum likelihood estimation of Beta regression parameters satisfies standard asymptotic theory. From **Lemma 1** (MLE convergence for Beta regression), with $d = 2p$ parameters (where p is the number of features), the estimation error is $O(d/n) = O(1/n)$. Combined with the approximation error $\delta_{\mathcal{F}_{\text{Beta}}}(p)$, we obtain the bias-variance decomposition:

$$\mathbb{E}[D_{\text{KL}}(p\|\hat{p}_{\text{Beta}})] = \underbrace{\delta_{\mathcal{F}_{\text{Beta}}}(p)}_{\text{approximation}} + \underbrace{O(1/n)}_{\text{estimation}}.$$

(ii) Piecewise uniform: From **Lemma 3** (histogram sample complexity), the KL divergence for a k-bin histogram estimated from n samples is bounded by the sum of approximation and estimation errors. The approximation term $O(1/k^2)$ follows from Part A(i) and Lemma 2. The estimation term $O(k \log k/n)$ arises from estimating k bin probabilities, where the $\log k$ factor accounts for regularization to prevent boundary issues.

Part C: Method Comparison. **(i) Beta preferred (unimodal case):** When $\delta_{\mathcal{F}_{\text{Beta}}}(p) \ll 1/k^2$ (Beta-like densities), we compare total errors:

$$\text{Beta:} \quad \delta_{\mathcal{F}_{\text{Beta}}}(p) + O(1/n)$$
$$\text{Piecewise:} \quad O(1/k^2) + O(k \log k/n)$$

For moderate $k \leq 9$ and $n \in [100, 10000]$, the piecewise method suffers from large estimation variance $O(k \log k/n)$ that dominates its approximation advantage, making Beta preferred.

(ii) Piecewise preferred (multimodal case): When $\delta_{\mathcal{F}_{\text{Beta}}}(p) \geq C_m > 0$ (multimodal), Beta incurs irreducible error from Lemma 4. For sufficiently large k and $n \geq Ck^2$ (ensuring $k \log k/n \ll 1/k^2$), the piecewise approximation error $O(1/k^2) \to 0$ dominates Beta's fixed bias, making piecewise preferred.

(iii) Asymptotic regime: As $k, n \to \infty$ with $k^2 \ll n$, the estimation error $k \log k/n \to 0$ faster than the approximation error $1/k^2 \to 0$. If $\delta_{\mathcal{F}_{\text{Beta}}}(p) > 0$ (non-Beta-like), then piecewise is asymptotically superior because:

$$\lim_{k,n\to\infty} \frac{\text{Piecewise error}}{\text{Beta error}} = \frac{0}{\delta_{\mathcal{F}_{\text{Beta}}}(p)} = 0$$

□

B.1 Technical Notes on Theorem 1

Regularization for finite samples. The piecewise uniform estimator in Lemma 3 uses Laplace-smoothed bin probabilities $\hat{p}_i = (n_i + 1)/(n + k)$ rather than raw empirical frequencies n_i/n to ensure finite KL divergence when bins may be empty for finite n. The added smoothing contributes $O(k/n)$ additional bias, which is dominated by the $O(k \log k/n)$ estimation variance for $n \gg k$.

Misspecification in Beta regression. Part B(i) applies when p can be well-approximated by the Beta family (small $\delta_{F_{\text{Beta}}}(p)$). Under general misspecification, the estimation error is $O(\kappa d/n)$ where $\kappa = \text{tr}(H^{-1}J)/d$ is the sandwich correction factor from the misspecified MLE literature (White, 1982). For moderate misspecification, $\kappa = O(1)$, preserving the $O(d/n)$ rate. The approximation error $\delta_{F_{\text{Beta}}}(p)$ dominates when p is far from any Beta distribution.

Regime condition. Part C(ii) requires $n \geq Ck^3$ (not Ck^2) to ensure the piecewise estimation error $O(k/n)$ is dominated by its approximation error $O(1/k^2)$.

B.2 Extension to Conditional Regression

Theorem 1 analyzes unconditional distributions $p(y)$ to characterize the fundamental approximation-estimation trade-off between parametric and non-parametric families. For conditional regression $p(y|x)$ with d-dimensional features:

Parameter counts:

- Beta regression: $2d$ parameters (weight vector $w \in \mathbb{R}^d$ for mean, scalar precision ϕ)
- Multi-threshold: kd parameters (weight vectors $w_i \in \mathbb{R}^d$ for each threshold τ_i)

Modified convergence rates: Under standard assumptions (bounded features $\|x\| \leq R$, Lipschitz smoothness of $p(y|x)$ in x), the estimation errors become:

- Beta: $E[D_{\text{KL}}(p\|\hat{p}_{\text{Beta}})] = \delta_{\mathcal{F}_{\text{Beta}}}(p) + O(d/n)$
- Multi-threshold: $E[D_{\text{KL}}(p\|\hat{q}_k)] \leq C\|p''\|_\infty^2/k^2 + O(kd\log(kd)/n)$

The approximation errors (Part A) remain unchanged, applying pointwise for each x. Part C's regime shifts from $n \in [100, 10000]$ and $k \leq 9$ to: Beta preferred when $n \gtrsim kd$ and the conditional distribution is approximately Beta-like at most x values. For our experiments with $d \approx 15$–30 and $k = 5$, this requires $n \gtrsim 75$–150, well within our training regime ($n \approx 500$–1000).

C Evaluation Metrics

We provide additional details about our choice of evaluation metrics, which are intended to evaluate the quality of confidence estimates for continuous correctness. They assess different aspects of model performance: discriminative ability, calibration quality, effectiveness of uncertainty quantification, and distributional fit.

Mean Absolute Error (MAE). Mean Absolute Error measures point prediction accuracy by computing the average absolute difference between predicted means and true correctness values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

where y_i is the true correctness score and $\hat{y}_i$ is the predicted mean for sample i. MAE provides an intuitive measure of prediction error on the original scale and is robust to distributional assumptions. Lower values indicate better performance, with MAE = 0 representing perfect predictions.

Continuous Expected Calibration Error (Continuous-ECE). Traditional Expected Calibration Error (Guo et al., 2017) measures calibration for binary classification by binning predicted probabilities and comparing them to empirical accuracy within each bin. We adapt this naturally to the continuous correctness setting. Let $\hat{\mu}_i$ denote the predicted mean correctness and $y_i \in [0, 1]$ denote the true correctness score for sample i. We partition predictions into K bins $\{B_1, \ldots, B_K\}$ based on their predicted mean values and compute:

$$\text{Continuous-ECE} = \sum_{k=1}^{K}\frac{|B_k|}{n}\left|\frac{1}{|B_k|}\sum_{i\in B_k}\hat{\mu}_i - \frac{1}{|B_k|}\sum_{i\in B_k}y_i\right|,$$

where n is the total number of samples and $|B_k|$ is the number of samples in bin k. This metric measures whether predicted mean correctness aligns with actual mean correctness across different confidence levels. A well-calibrated model should have Continuous-ECE ≈ 0, indicating that when the model predicts a mean correctness of, say, 0.7, the actual average correctness of such predictions is also close to 0.7. We use $K = 10$ bins for all experiments.

Interval Score (IS). The Interval Score evaluates prediction interval quality by balancing sharpness (interval width) and coverage (whether true values fall within intervals). For a $(1 - \alpha)$ prediction interval $[\ell_i, u_i]$, the IS for sample i is:

$$\text{IS}_i = (u_i - \ell_i) + \frac{2}{\alpha}(\ell_i - y_i)\mathbb{1}(y_i < \ell_i) + \frac{2}{\alpha}(y_i - u_i)\mathbb{1}(y_i > u_i),$$

where $\mathbb{1}(\cdot)$ is the indicator function. The first term penalizes wide intervals (poor sharpness), while the second and third terms heavily penalize coverage violations. The mean interval score across all samples is reported, with lower values indicating better interval quality. For $\alpha = 0.1$ (90% intervals), coverage violations are penalized at $20\times$ the magnitude of the violation (Winkler, 1972; Gneiting & Raftery, 2007).

Continuous Ranked Probability Score (CRPS). CRPS is a proper scoring rule that measures the quality of probabilistic forecasts by comparing the predicted cumulative distribution function (CDF) F to the empirical CDF of the observation y_i:

$$\text{CRPS}(F, y_i) = \int_{-\infty}^{\infty} [F(y) - \mathbb{1}(y \geq y_i)]^2 \, dy,$$

where $\mathbb{1}(y \geq y_i)$ represents the empirical CDF of the observation. CRPS generalizes Mean Absolute Error to full probability distributions: it equals MAE when F is a point mass. Lower values indicate better distributional predictions. CRPS simultaneously rewards accuracy, sharpness, and calibration, making it a comprehensive measure of distributional forecast quality. We report the mean CRPS across all test samples (Gneiting & Raftery, 2007).

Selective Mean Absolute Error (Selective-MAE). Selective-MAE evaluates whether uncertainty estimates enable effective selective prediction. For retention rate $\rho \in (0, 1]$, we compute MAE on the $\rho \times n$ most reliable predictions as ranked by lowest predicted uncertainty:

$$\text{Selective-MAE}(\rho) = \frac{1}{|\mathcal{S}_\rho|} \sum_{i \in \mathcal{S}_\rho} |y_i - \hat{y}_i|,$$

where $\mathcal{S}_\rho$ contains the samples ranked by uncertainty. For methods with explicit uncertainty estimates, we select based on low predicted uncertainty; we do not report this metric for point prediction methods. Lower Selective-MAE at reduced retention indicates successful identification of reliable predictions. We report Selective-MAE at $\rho = 0.5$, measuring error on the 50% most confident predictions (El-Yaniv & Wiener, 2010).

Metrics that are leveraged in Appendices (and not in the main file) follow:

Pearson Correlation (Pearson-r). Pearson correlation measures the linear association between predicted and true correctness values:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}},$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the mean true and predicted values, respectively. Pearson-r ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with 0 indicating no linear relationship. Unlike MAE or R^2, Pearson-r is scale-invariant and only captures linear association, not prediction accuracy. Higher positive values indicate stronger agreement between predictions and observations, though perfect correlation ($r = 1$) can occur even with systematically biased predictions.

Coefficient of Determination (R^2). R^2 quantifies the proportion of variance in correctness scores explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}},$$

where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares. $R^2 = 1$ indicates perfect predictions, $R^2 = 0$ indicates performance equivalent to predicting the mean, and $R^2 < 0$ indicates predictions worse than the mean baseline. While R^2 provides an interpretable measure of explained variance, it can be negative for models that do not minimize squared error during training (e.g., quantile regression), and does not directly assess distributional quality or calibration.

Negative Log-Likelihood (NLL). NLL measures how well the predicted probability distribution assigns likelihood to the observed correctness values:

$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log p(y_i | \theta_i),$$

where $p(y_i|\theta_i)$ is the probability density (or mass) assigned to observation y_i under the predicted distribution with parameters θ_i. Lower NLL indicates better probabilistic predictions, with the

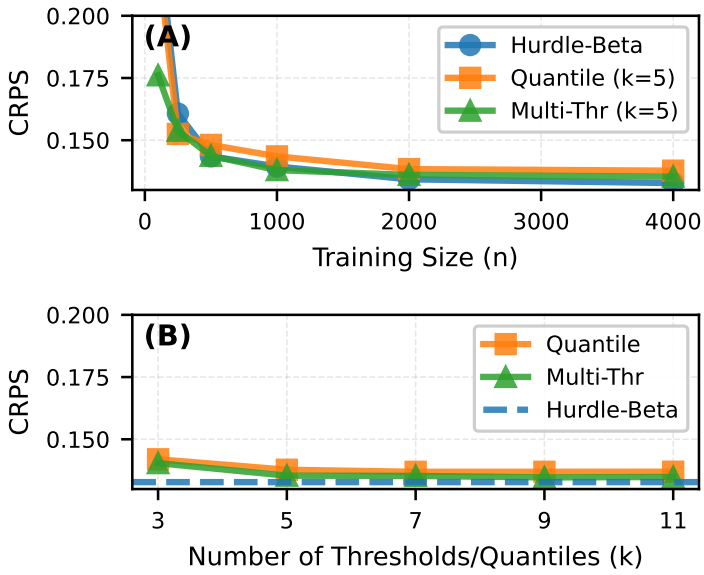


Figure 4: Ablation study including three distributional methods on HotpotQA, evaluated via CRPS. **(A)** Effect of training size (n) with $k = 5$ thresholds fixed. **(B)** Effect of number of thresholds (k) with full training data. Hurdle Beta does not depend on k.

model assigning higher likelihood to observed values. NLL is a proper scoring rule that strongly penalizes confident wrong predictions. For Beta regression, $p(y|\theta)$ is the Beta density; for Quantile Regression, we approximate the density using piecewise uniform distributions between predicted quantiles. Note that NLL values are not directly comparable across fully continuous distributions and mixture distributions, as they operate under different probability measures.

D Additional Experimental Results

D.1 Ablation: Training Size and Thresholds

To understand how training data size and model complexity for piecewise uniform methods affect performance, we conduct ablation studies on HotpotQA with generations from Llama-3.3-70B at temperature 0.0. Figure 4 shows CRPS as training size (n) varies from 100 to 4000 samples (Panel A, with $k = 5$ fixed), and as the number of thresholds (k) varies from 3 to 11 (Panel B, with full training data). Panel A demonstrates that all methods benefit from more data. In this case, all methods converge by $n \approx 1000$, suggesting that distributional modeling benefits plateau with moderate training data on this task. Panel B shows diminishing returns beyond $k = 5$ thresholds, with performance essentially flat for $k \geq 5$, indicating that simpler piecewise approximations are sufficient in this case.

D.2 Extended Results

Here we provide detailed results across several metrics for all 7 datasets. We also include additional metrics that were not considered in the main text. There is a separate table for each dataset. Recall that not all methods can compute all metrics. NLL for Hurdle Beta is not comparable with other distributional methods since it is a mixture distribution and not fully continuous, and is therefore not shown.

Table 4: Confidence estimation results for **HotpotQA**: 8 evaluation metrics and 2 LLMs using generations at temperature zero. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta in situations with no concentrated mass near perfect correctness. MT refers to multi-threshold.

	Method	MAE↓	ECE↓	IS@90%↓	CRPS↓	r↑	R^2↑	NLL↓	Sel-MAE@50%↓
Granite4.0-32B	Beta Reg	0.223±.010	0.095±.016	0.834±.025	0.157±.008	0.328±.060	-0.021±.093	-0.642±.141	0.232±.017
	Hurdle Beta	0.200±.008	0.027±.014	0.772±.006	0.129±.006	0.409±.049	0.159±.046	-	0.170±.010
	MT Binary Reg	0.205±.006	0.029±.002	0.781±.017	0.132±.004	0.389±.046	0.137±.038	0.518±.331	0.178±.012
	Quantile Reg	0.211±.006	0.046±.011	0.780±.039	0.136±.003	0.327±.052	0.058±.074	0.314±.228	0.195±.015
	Linear Reg	0.211±.006	0.038±.012	N/A	N/A	0.348±.051	0.107±.056	N/A	N/A
	Random Forest	0.189±.009	0.063±.022	1.034±.040	N/A	0.420±.077	0.112±.106	N/A	0.166±.012
	Cal-Logit	0.205±.006	0.032±.012	N/A	N/A	0.360±.074	0.118±.056	N/A	N/A
	Cal-Similarity	0.225±.009	0.021±.013	N/A	N/A	0.130±.045	0.007±.015	N/A	N/A
	Cal-Clustering	0.225±.009	0.020±.012	N/A	N/A	0.120±.053	0.005±.014	N/A	N/A
Llama3.3-70B	Beta Reg	0.219±.010	0.085±.012	0.824±.027	0.154±.006	0.355±.060	0.026±.069	-0.822±.143	0.226±.008
	Hurdle Beta	0.198±.007	0.029±.011	0.762±.023	0.128±.004	0.457±.039	0.204±.033	-	0.166±.008
	MT Binary Reg	0.206±.008	0.032±.014	0.779±.032	0.133±.003	0.423±.033	0.169±.025	0.599±.310	0.184±.009
	Quantile Reg	0.213±.010	0.040±.015	0.766±.031	0.136±.005	0.362±.054	0.100±.058	0.362±.276	0.201±.013
	Linear Reg	0.212±.010	0.037±.021	N/A	N/A	0.378±.044	0.134±.039	N/A	N/A
	Random Forest	0.200±.010	0.070±.012	1.018±.064	N/A	0.377±.026	0.064±.020	N/A	0.181±.009
	Cal-Logit	0.198±.010	0.030±.012	N/A	N/A	0.432±.052	0.176±.045	N/A	N/A
	Cal-Similarity	0.230±.008	0.016±.012	N/A	N/A	0.169±.052	0.023±.018	N/A	N/A
	Cal-Clustering	0.233±.009	0.015±.013	N/A	N/A	0.070±.027	0.000±.011	N/A	N/A

Table 5: Confidence estimation results for **NaturalQuestions**: 8 evaluation metrics and 2 LLMs using generations at temperature zero. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta in situations with no concentrated mass near perfect correctness. MT refers to multi-threshold.

	Method	MAE↓	ECE↓	IS@90%↓	CRPS↓	r↑	R^2↑	NLL↓	Sel-MAE@50%↓
Granite4.0-32B	Beta Reg	0.211±.011	0.091±.022	0.841±.031	0.152±.008	0.323±.061	-0.026±.108	-0.529±.151	0.227±.010
	Hurdle Beta	0.189±.011	0.033±.013	0.744±.009	0.124±.008	0.411±.027	0.164±.021	-	0.161±.014
	MT Binary Reg	0.191±.012	0.035±.014	0.766±.013	0.126±.006	0.392±.027	0.142±.030	0.499±.418	0.161±.015
	Quantile Reg	0.197±.009	0.054±.017	0.784±.055	0.130±.005	0.305±.036	0.055±.048	0.318±.248	0.172±.009
	Linear Reg	0.197±.011	0.035±.014	N/A	N/A	0.326±.043	0.093±.046	N/A	N/A
	Random Forest	0.192±.006	0.068±.011	0.982±.072	N/A	0.342±.039	0.022±.055	N/A	0.168±.009
	Cal-Logit	0.188±.013	0.025±.006	N/A	N/A	0.392±.053	0.148±.043	N/A	N/A
	Cal-Similarity	0.210±.014	0.022±.011	N/A	N/A	0.142±.060	0.009±.029	N/A	N/A
	Cal-Clustering	0.210±.014	0.020±.012	N/A	N/A	0.093±.055	-0.001±.019	N/A	N/A
Llama3.3-70B	Beta Reg	0.220±.007	0.085±.021	0.827±.008	0.156±.007	0.318±.049	-0.018±.069	-0.695±.160	0.236±.007
	Hurdle Beta	0.202±.005	0.027±.008	0.754±.021	0.130±.004	0.382±.022	0.137±.020	-	0.179±.008
	MT Binary Reg	0.204±.004	0.033±.013	0.756±.022	0.131±.003	0.380±.029	0.130±.024	0.586±.409	0.178±.012
	Quantile Reg	0.210±.007	0.035±.016	0.766±.036	0.135±.004	0.310±.035	0.069±.038	0.187±.355	0.202±.010
	Linear Reg	0.211±.007	0.027±.004	N/A	N/A	0.329±.038	0.088±.042	N/A	N/A
	Random Forest	0.210±.007	0.085±.017	1.154±.149	N/A	0.297±.055	-0.047±.079	N/A	0.194±.012
	Cal-Logit	0.203±.005	0.024±.005	N/A	N/A	0.367±.015	0.128±.013	N/A	N/A
	Cal-Similarity	0.224±.007	0.019±.013	N/A	N/A	0.040±.048	-0.007±.011	N/A	N/A
	Cal-Clustering	0.224±.007	0.019±.015	N/A	N/A	0.000±.000	-0.008±.009	N/A	N/A

Table 6: Confidence estimation results for **TriviaQA**: 8 evaluation metrics and 2 LLMs using generations at temperature zero. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta in situations with no concentrated mass near perfect correctness. MT refers to multi-threshold.

	Method	MAE↓	ECE↓	IS@90%↓	CRPS↓	r↑	R^2↑	NLL↓	Sel-MAE@50%↓
Granite4.0-32B	Beta Reg	0.165±.012	0.057±.012	0.657±.021	0.105±.013	0.455±.089	0.175±.097	-2.287±.265	0.132±.014
	Hurdle Beta	0.151±.010	0.033±.013	0.750±.050	0.094±.010	0.547±.067	0.285±.068	-	0.118±.012
	MT Binary Reg	0.168±.009	0.052±.015	0.726±.012	0.103±.009	0.549±.079	0.262±.060	1.723±.604	0.146±.007
	Quantile Reg	0.142±.016	0.087±.018	0.730±.098	0.100±.010	0.489±.089	0.061±.176	0.320±1.218	0.102±.020
	Linear Reg	0.164±.012	0.040±.014	N/A	N/A	0.502±.085	0.228±.090	N/A	N/A
	Random Forest	0.148±.013	0.054±.018	0.862±.153	N/A	0.490±.058	0.173±.089	N/A	0.100±.015
	Cal-Logit	0.155±.010	0.027±.010	N/A	N/A	0.488±.041	0.231±.044	N/A	N/A
	Cal-Similarity	0.192±.006	0.023±.009	N/A	N/A	0.229±.059	0.043±.029	N/A	N/A
	Cal-Clustering	0.195±.006	0.020±.006	N/A	N/A	0.180±.091	0.027±.033	N/A	N/A
Llama3.3-70B	Beta Reg	0.133±.004	0.023±.009	0.622±.055	0.081±.004	0.348±.087	0.117±.058	-2.836±.126	0.116±.012
	Hurdle Beta	0.129±.004	0.015±.005	0.733±.027	0.079±.004	0.397±.111	0.161±.087	-	0.111±.009
	MT Binary Reg	0.149±.004	0.045±.008	0.701±.040	0.091±.003	0.396±.111	0.121±.086	1.806±.828	0.135±.004
	Quantile Reg	0.102±.006	0.084±.007	0.721±.110	0.079±.004	0.373±.118	-0.057±.110	-0.839±1.125	0.082±.019
	Linear Reg	0.134±.006	0.020±.008	N/A	N/A	0.375±.097	0.116±.079	N/A	N/A
	Random Forest	0.127±.005	0.050±.010	0.793±.075	N/A	0.360±.105	0.036±.144	N/A	0.081±.011
	Cal-Logit	0.126±.004	0.020±.006	N/A	N/A	0.415±.089	0.165±.083	N/A	N/A
	Cal-Similarity	0.149±.007	0.007±.006	N/A	N/A	0.081±.088	0.010±.015	N/A	N/A
	Cal-Clustering	0.149±.007	0.008±.006	N/A	N/A	0.075±.089	0.006±.012	N/A	N/A

Table 7: Confidence estimation results for **CNN/DailyMail**: 8 evaluation metrics and 2 LLMs using generations at temperature zero. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta in situations with no concentrated mass near perfect correctness. MT refers to multi-threshold.

	Method	MAE↓	ECE↓	IS@90%↓	CRPS↓	r↑	R^2↑	NLL↓	Sel-MAE@50%↓
Granite4.0-32B	Beta Reg	0.042±.002	0.004±.003	0.223±.013	0.030±.001	0.300±.076	0.072±.070	-1.518±.045	0.044±.003
	Hurdle Beta	-	-	-	-	-	-	-	-
	MT Binary Reg	0.043±.002	0.006±.005	0.326±.015	0.033±.002	0.215±.044	0.022±.054	-1.219±.076	0.041±.003
	Quantile Reg	0.042±.002	0.004±.004	0.251±.012	0.032±.001	0.298±.079	0.064±.081	-1.098±.065	0.040±.003
	Linear Reg	0.042±.002	0.004±.003	N/A	N/A	0.299±.076	0.072±.070	N/A	N/A
	Random Forest	0.043±.002	0.008±.006	0.247±.015	N/A	0.298±.067	0.030±.111	N/A	0.041±.003
	Cal-Logit	0.043±.002	0.004±.003	N/A	N/A	0.188±.078	0.019±.045	N/A	N/A
	Cal-Similarity	0.043±.002	0.004±.004	N/A	N/A	0.117±.069	0.001±.033	N/A	N/A
	Cal-Clustering	0.044±.003	0.003±.003	N/A	N/A	0.000±.000	-0.006±.009	N/A	N/A
Llama3.3-70B	Beta Reg	0.049±.004	0.005±.002	0.259±.016	0.035±.002	0.304±.082	0.086±.057	-1.363±.067	0.054±.005
	Hurdle Beta	-	-	-	-	-	-	-	-
	MT Binary Reg	0.050±.002	0.006±.004	0.383±.011	0.037±.002	0.267±.060	0.046±.077	-1.077±.076	0.046±.003
	Quantile Reg	0.049±.004	0.005±.003	0.277±.025	0.037±.002	0.291±.088	0.074±.060	-0.989±.090	0.044±.003
	Linear Reg	0.049±.004	0.005±.001	N/A	N/A	0.303±.081	0.085±.055	N/A	N/A
	Random Forest	0.053±.003	0.016±.005	0.330±.029	N/A	0.167±.055	-0.096±.071	N/A	0.052±.003
	Cal-Logit	0.050±.005	0.006±.004	N/A	N/A	0.227±.124	0.034±.094	N/A	N/A
	Cal-Similarity	0.051±.004	0.003±.001	N/A	N/A	0.050±.055	-0.002±.012	N/A	N/A
	Cal-Clustering	0.051±.004	0.003±.001	N/A	N/A	0.000±.000	-0.002±.001	N/A	N/A

Table 8: Confidence estimation results for **XSum**: 8 evaluation metrics and 2 LLMs using generations at temperature zero. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta in situations with no concentrated mass near perfect correctness. MT refers to multi-threshold.

	Method	MAE↓	ECE↓	IS@90%↓	CRPS↓	r↑	R^2↑	NLL↓	Sel-MAE@50%↓
Granite4.0-32B	Beta Reg	0.054±.003	0.006±.003	0.316±.030	0.039±.002	0.229±.033	0.027±.053	-1.215±.074	0.057±.003
	Hurdle Beta	-	-	-	-	-	-	-	-
	MT Binary Reg	0.057±.003	0.014±.006	0.495±.011	0.043±.002	0.241±.082	-0.078±.182	-0.973±.071	0.050±.003
	Quantile Reg	0.054±.002	0.006±.002	0.338±.038	0.041±.002	0.237±.024	0.025±.043	-0.950±.093	0.048±.002
	Linear Reg	0.055±.003	0.005±.003	N/A	N/A	0.220±.033	0.014±.062	N/A	N/A
	Random Forest	0.057±.003	0.012±.005	0.355±.039	N/A	0.216±.066	-0.044±.077	N/A	0.053±.003
	Cal-Logit	0.055±.003	0.003±.002	N/A	N/A	0.038±.054	-0.005±.011	N/A	N/A
	Cal-Similarity	0.055±.003	0.003±.002	N/A	N/A	0.082±.093	0.001±.030	N/A	N/A
	Cal-Clustering	0.055±.003	0.003±.002	N/A	N/A	-0.003±.142	-0.009±.022	N/A	N/A
Llama3.3-70B	Beta Reg	0.053±.002	0.004±.003	0.296±.037	0.038±.002	0.183±.055	0.013±.032	-1.259±.083	0.051±.003
	Hurdle Beta	-	-	-	-	-	-	-	-
	MT Binary Reg	0.056±.004	0.013±.006	0.489±.011	0.042±.002	0.216±.069	-0.089±.139	-0.933±.186	0.050±.003
	Quantile Reg	0.053±.002	0.005±.003	0.318±.036	0.039±.002	0.188±.044	0.009±.040	-0.963±.055	0.048±.004
	Linear Reg	0.053±.002	0.004±.004	N/A	N/A	0.180±.054	0.005±.045	N/A	N/A
	Random Forest	0.055±.002	0.015±.004	0.368±.031	N/A	0.247±.030	-0.069±.073	N/A	0.053±.001
	Cal-Logit	0.053±.002	0.002±.001	N/A	N/A	0.104±.081	0.002±.031	N/A	N/A
	Cal-Similarity	0.053±.003	0.001±.001	N/A	N/A	0.043±.088	-0.002±.015	N/A	N/A
	Cal-Clustering	0.053±.003	0.001±.001	N/A	N/A	0.000±.000	0.000±.001	N/A	N/A

Table 9: Confidence estimation results for **Biographies**: 8 evaluation metrics and 2 LLMs using generations at temperature zero. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta in situations with no concentrated mass near perfect correctness. MT refers to multi-threshold.

	Method	MAE↓	ECE↓	IS@90%↓	CRPS↓	r↑	R^2↑	NLL↓	Sel-MAE@50%↓
Granite4.0-32B	Beta Reg	0.047±.008	0.017±.009	0.387±.141	0.037±.008	0.378±.215	-0.537±1.429	-0.785±1.066	0.046±.019
	Hurdle Beta	-	-	-	-	-	-	-	-
	MT Binary Reg	0.050±.007	0.024±.019	0.422±.022	0.044±.003	0.529±.144	-0.093±.338	-0.925±.133	0.042±.007
	Quantile Reg	0.047±.010	0.020±.010	0.486±.298	0.038±.008	0.359±.231	-0.888±2.229	-0.974±.152	0.041±.026
	Linear Reg	0.048±.010	0.020±.008	N/A	N/A	0.376±.233	-0.940±2.347	N/A	N/A
	Random Forest	0.043±.003	0.013±.008	0.293±.093	N/A	0.458±.326	0.206±.316	N/A	0.034±.011
	Cal-Logit	0.047±.005	0.009±.003	N/A	N/A	0.080±.175	-0.047±.110	N/A	N/A
	Cal-Similarity	0.047±.005	0.009±.004	N/A	N/A	-0.036±.063	-0.039±.012	N/A	N/A
	Cal-Clustering	0.047±.005	0.008±.005	N/A	N/A	0.013±.057	-0.021±.016	N/A	N/A
Llama3.3-70B	Beta Reg	0.045±.007	0.020±.013	0.264±.061	0.032±.005	0.483±.268	0.156±.304	-1.409±.225	0.038±.006
	Hurdle Beta	-	-	-	-	-	-	-	-
	MT Binary Reg	0.059±.004	0.046±.008	0.472±.023	0.054±.001	0.443±.347	-0.589±.931	-0.463±.482	0.062±.010
	Quantile Reg	0.047±.008	0.019±.010	0.302±.092	0.035±.006	0.399±.314	0.012±.349	-1.041±.367	0.041±.006
	Linear Reg	0.045±.007	0.020±.014	N/A	N/A	0.478±.266	0.139±.309	N/A	N/A
	Random Forest	0.048±.004	0.028±.006	0.366±.095	N/A	0.354±.386	-0.167±.706	N/A	0.049±.006
	Cal-Logit	0.051±.012	0.021±.008	N/A	N/A	0.192±.106	-0.131±.217	N/A	N/A
	Cal-Similarity	0.051±.012	0.017±.008	N/A	N/A	-0.047±.051	-0.131±.183	N/A	N/A
	Cal-Clustering	0.051±.012	0.017±.008	N/A	N/A	-0.023±.046	-0.127±.184	N/A	N/A

Table 10: Confidence estimation results for **BigCodeBench**: 8 evaluation metrics and 2 LLMs using generations at temperature zero. Mean ± std shown over 5 runs, each with a random 80%-20% train-test split. Hurdle-Beta is identical to Beta in situations with no concentrated mass near perfect correctness. MT refers to multi-threshold.

	Method	MAE↓	ECE↓	IS@90%↓	CRPS↓	r↑	R^2↑	NLL↓	Sel-MAE@50%↓
Qwen2.5-Coder-7B	Beta Reg	0.375±.017	0.047±.012	1.000±.000	0.221±.005	0.237±.025	0.046±.013	-1.829±.213	0.366±.024
	Hurdle Beta	0.371±.018	0.062±.023	1.063±.011	0.223±.005	0.232±.041	0.042±.024	-	0.355±.022
	MT Binary Reg	0.365±.015	0.037±.012	1.178±.035	0.225±.006	0.243±.024	0.057±.013	-0.426±.075	0.346±.020
	Quantile Reg	0.354±.011	0.132±.016	1.000±.000	0.222±.007	0.217±.018	-0.062±.028	-5.123±.776	0.346±.013
	Linear Reg	0.362±.015	0.038±.007	N/A	N/A	0.242±.026	0.054±.015	N/A	N/A
	Random Forest	0.364±.019	0.077±.025	1.011±.031	N/A	0.215±.061	0.013±.051	N/A	0.338±.025
	Cal-Logit	0.383±.017	0.021±.012	N/A	N/A	0.049±.053	0.000±.007	N/A	N/A
	Cal-Similarity	0.366±.019	0.034±.014	N/A	N/A	0.214±.067	0.040±.036	N/A	N/A
	Cal-Clustering	0.377±.015	0.027±.011	N/A	N/A	0.147±.025	0.017±.005	N/A	N/A
Ministral-8B	Beta Reg	0.360±.013	0.052±.020	1.000±.000	0.212±.007	0.264±.063	0.051±.016	-2.176±.144	0.348±.020
	Hurdle Beta	0.355±.015	0.049±.022	1.075±.005	0.213±.009	0.267±.070	0.057±.029	-	0.333±.022
	MT Binary Reg	0.350±.014	0.031±.012	1.164±.016	0.216±.010	0.262±.061	0.065±.026	-0.495±.105	0.331±.018
	Quantile Reg	0.329±.020	0.144±.023	1.001±.003	0.211±.010	0.254±.063	-0.069±.071	-4.095±.891	0.341±.026
	Linear Reg	0.345±.015	0.034±.010	N/A	N/A	0.265±.064	0.067±.034	N/A	N/A
	Random Forest	0.351±.011	0.057±.010	1.021±.013	N/A	0.216±.037	0.017±.027	N/A	0.322±.016
	Cal-Logit	0.366±.012	0.025±.009	N/A	N/A	0.074±.053	0.000±.012	N/A	N/A
	Cal-Similarity	0.359±.014	0.030±.014	N/A	N/A	0.162±.072	0.021±.026	N/A	N/A
	Cal-Clustering	0.368±.010	0.021±.010	N/A	N/A	0.019±.049	-0.003±.002	N/A	N/A