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.

PROBABILISTIC EVALUATION OF LARGE LANGUAGE MODELS WITH AMORTIZED BAYESIAN RELIABILITY ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) evaluation increasingly relies on automated performance metrics to compare models across tasks, yet the reliability and uncertainty of these evaluations remain insufficiently characterized. Conventional evaluation metrics typically provide point estimates of model performance, making it difficult to determine whether an observed performance difference reflects a meaningful improvement or uncertainty in the evaluation process. In this work, we propose a Bayesian inference-based performance metric for quantifying the reliability of LLM evaluation. The proposed approach models evaluation outcomes probabilistically, enabling performance estimates to be accompanied by uncertainty information and providing a principled basis for assessing the confidence and reliability of model comparisons. We evaluate the proposed metric across multiple language models and benchmarks, including PolyGemma, BERT, and Polyglot, and compare its behavior against established evaluation approaches. Our experimental analysis demonstrates how Bayesian inference can expose uncertainty that is otherwise hidden by conventional point-based metrics and provide a more informative characterization of model performance. The results highlight the importance of accounting for uncertainty when ranking or comparing LLMs and suggest that reliable evaluation should consider not only the magnitude of measured performance but also the statistical confidence associated with that measurement. The proposed framework provides a general and interpretable approach for uncertainty-aware evaluation of LLMs and can be applied to systematic comparisons across models, datasets, and evaluation settings.We formulate LLM evaluation as Bayesian inference over a continuous latent semantic-quality variable, using a Gaussian prior to represent prior beliefs and uncertainty over graded semantic relatedness.

Keywords: Large Language Models (LLMs), Bayesian Inference, Uncertainty Quantification, LLM Evaluation, Performance Reliability, Probabilistic Evaluation, Model Comparison, Robust Evaluation.

1 INTRODUCTION

Evaluation has been a fundamental driver of progress in Artificial Intelligence (AI), providing systematic means to measure, compare, and validate the capabilities of intelligent systems. As AI has evolved, evaluation methodologies have progressed from well-defined, programmable metrics to more comprehensive frameworks capable of assessing complex and open-ended forms of intelligence. Early machine learning systems were primarily evaluated using metrics such as accuracy, precision, recall, and error measures for structured tasks including classification and regression, while the emergence of deep learning introduced increasingly complex models and applications that demanded more sophisticated evaluation standards. The Turing Test represented an important shift toward assessing human-like intelligent behavior through interaction, and the recent emergence of Large Language Models (LLMs) and generative AI has further transformed the evaluation landscape. LLMs demonstrate remarkable generalization, adaptability, and transfer capabilities across diverse tasks and domains, but their open-ended and context-dependent outputs expose significant

limitations in conventional evaluation metrics. For instance, metrics such as BLEU and ROUGE may measure lexical similarity but often fail to adequately capture semantic correctness, factuality, logical reasoning, coherence, relevance, and creativity. Consequently, modern AI evaluation must extend beyond task-specific performance to assess broader properties such as robustness, fairness, interpretability, reliability, safety, and adaptability in realistic and dynamic environments. Although human evaluation can provide valuable insights into these dimensions, human annotation is often subjective, expensive, time-consuming, and difficult to scale. This motivates my research interest in developing scalable, reliable, and human-aligned evaluation frameworks for modern AI systems that integrate automated quantitative assessment, model-based evaluation, human feedback, and principled statistical or causal analysis. In particular, I aim to investigate methodologies that can determine not only whether an AI system produces a correct response, but also why it succeeds or fails, how consistently its capabilities generalize across diverse settings, and whether its observed performance reflects genuine reasoning or reliance on undesirable shortcuts. Such evaluation frameworks can contribute to the development of AI systems that are not only more capable, but also reliable, interpretable, robust, and trustworthy.

1.1 Motivation

Papers to be submitted to ICLR 2027 must be prepared according to the instructions presented here.

Authors are required to use the ICLR LaTeX style files obtainable at the ICLR website. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

1.2 Contribution

- Bayesian Evaluation Metric: We propose a Bayesian inference-based performance metric that explicitly incorporates uncertainty into LLM evaluation, moving beyond conventional point-based performance estimates.
- Reliability-Aware Model Comparison: We introduce a probabilistic framework for quantifying the confidence, stability, and reliability of observed model performance, enabling more informed comparisons between LLMs.
- Comprehensive Empirical Evaluation: We evaluate the proposed approach across diverse language models, including PolyGemma, BERT, and Polyglot, and compare it with state-of-the-art evaluation approaches to demonstrate the benefits of uncertainty-aware LLM evaluation.

.

2 Related Work

Recent advances in large language models (LLMs) have expanded their use from conventional language generation to multimodal reasoning, causal inference, continual learning, foundation models, and multilingual learning. However, the increasing capabilities of LLMs have also created a need for systematic evaluation frameworks that can assess not only task-level performance but also reasoning ability, generalization, robustness, knowledge acquisition, and the quality of model-generated data. The following studies provide relevant perspectives for understanding the current landscape of LLM capabilities and evaluation.

2.1 LLMs for Multimodal Understanding and Generation

The integration of LLMs with visual models represents an important direction for evaluating the ability of language models to operate beyond text-only environments. **Paper 2** proposes a framework that connects frozen text-only LLMs with pretrained image encoders, decoders, and text-to-image generation models by learning mappings between their embedding spaces. The resulting system supports multiple multimodal capabilities, including image retrieval, image generation, and multimodal dialogue. An important aspect of this work is its ability to process arbitrarily interleaved image and text inputs and generate either textual or visual outputs. The study evaluates the

proposed model against existing generation approaches and demonstrates improved performance for tasks requiring complex and context-dependent language.

From an LLM evaluation perspective, **Paper 2** highlights that conventional text-only benchmarks are insufficient for characterizing modern LLM capabilities. Evaluation must increasingly consider **cross-modal understanding, contextual grounding, retrieval-versus-generation decisions, and consistency between linguistic representations and visual outputs**. In particular, the ability of a model to determine whether an image should be retrieved or generated introduces an additional decision-making dimension that cannot be captured through standard language-generation metrics alone. Thus, multimodal LLMs motivate evaluation protocols that jointly assess language, vision, grounding, and generation quality.

2.2 Causal Reasoning and Intervention-Based Evaluation

Another important dimension of LLM evaluation is the assessment of **causal reasoning**, which goes beyond statistical association and requires understanding relationships between interventions, variables, and outcomes. **Paper 1** investigates causal effects in complex spatiotemporal systems by introducing a framework for estimating Propagated Causal Effects (PCE). The approach exploits physical system topology to constrain probabilistic inference and evaluates causal relationships through statistical hypothesis testing. The framework is demonstrated across linear, nonlinear, non-Markovian, and multilayer network settings.

Although the framework is not itself an LLM-based method, it provides an important methodological perspective for evaluating AI systems that claim causal reasoning capabilities. In particular, the distinction between an inference schedule and an actual causal mechanism is critical. **Paper 1** demonstrates that structural assumptions used to improve computational efficiency should not automatically be interpreted as evidence of causality. This observation is highly relevant to LLM evaluation because LLMs can generate plausible causal explanations without necessarily possessing genuine causal understanding. Consequently, evaluating LLMs for causal reasoning should involve **intervention-based tests, counterfactual scenarios, structural consistency, and robustness to changes in causal assumptions**, rather than relying solely on linguistic plausibility or benchmark accuracy.

Similarly, **Paper 3** directly addresses causal discovery using continuous constrained optimization and neural architectures. The proposed approach incorporates interventional data into causal DAG learning and employs expressive neural models such as normalizing flows. Its evaluation considers both perfect and imperfect interventions, including cases where the intervention targets are unknown. The study demonstrates that incorporating intervention information can improve causal structure learning and alleviate identifiability problems.

For LLM evaluation, the implications of **Paper 3** are significant. It suggests that evaluating reasoning systems solely using observational question-answering datasets may provide an incomplete picture of their causal capabilities. More rigorous evaluation should introduce **controlled interventions and counterfactual modifications** and examine whether model predictions change consistently with the underlying causal structure. Such evaluation can distinguish genuine causal reasoning from pattern matching or memorization.

2.3 Concept Learning, Knowledge Retention, and Continual Learning

Understanding how LLMs acquire and retain knowledge is another important component of model evaluation. **Paper 4** examines the acquisition, forgetting, interference, and synergy of individual concepts during continual pretraining. Rather than evaluating models only through their output behavior, the study investigates the internal concept circuits associated with particular concepts and analyzes their topology using graph-based metrics.

The findings indicate that concept circuits provide meaningful signals of learning and forgetting and exhibit characteristic temporal patterns during continual training. The study further shows that concepts with larger learning gains can experience greater subsequent forgetting, while semantically related concepts may produce stronger interference. Conversely, some concepts facilitate the learning of others.

These findings broaden the scope of LLM evaluation from **behavioral evaluation to mechanistic and representation-level evaluation**. Conventional benchmarks can reveal whether a model answers a question correctly, but they provide limited information about *how* the knowledge was represented or whether it remains stable after additional training. **Paper 4** therefore motivates evaluation dimensions such as **knowledge retention, catastrophic forgetting, concept interference, transferability, and internal representation stability**. For comprehensive LLM evaluation, these factors are particularly important in continual-learning and domain-adaptation scenarios.

2.4 Foundation Models and Evaluation of Causal Inference

The emergence of foundation-model paradigms has also influenced causal inference. **Paper 5** introduces CausalFM, a framework for training prior-data fitted networks (PFNs) as foundation models for causal inference. The approach constructs Bayesian priors from structural causal models and trains transformer-based models to perform in-context causal inference for settings including back-door, front-door, and instrumental-variable adjustment.

The importance of **Paper 5** for LLM evaluation lies in its demonstration that transformer-based foundation models can potentially learn generalizable reasoning patterns across multiple causal inference settings rather than being optimized for a single task. This raises an important evaluation question: **can a model generalize its reasoning capability across different causal structures and interventions, or has it merely learned task-specific patterns?**

Consequently, foundation-model evaluation should consider **cross-task generalization, in-context learning, structural generalization, and transfer to previously unseen causal settings**. Evaluating a model on only the same distribution used during training may overestimate its actual reasoning capability. The CausalFM framework illustrates the value of evaluating models across multiple causal inference scenarios and provides motivation for broader evaluation frameworks that test generalization beyond individual benchmark tasks.

2.5 Multilingual LLM Evaluation and Synthetic Data Quality

LLM evaluation becomes particularly challenging in multilingual settings because model performance can vary substantially across languages. **Paper 6** systematically studies the selection of multilingual teacher models for generating synthetic supervised fine-tuning data. The authors evaluate ten language models across six typologically diverse languages, generating more than 1.4 million synthetic examples and training hundreds of student models. Rather than assuming that the largest model is necessarily the best teacher, the study introduces the **POLYGLOT SCORE** to relate intrinsic properties of generated data to downstream student performance.

The results indicate that model scale alone is insufficient for predicting teacher effectiveness. Characteristics such as prompt diversity, response length, and fluency provide substantially more information about synthetic-data quality. The work therefore demonstrates that **LLM evaluation should not rely exclusively on model size or direct benchmark accuracy**. The quality of generated data and its downstream utility are also important indicators of model capability.

For multilingual LLMs, this observation is particularly relevant because a model may achieve strong aggregate benchmark performance while producing low-quality or culturally inappropriate outputs in less-resourced languages. **Paper 6** therefore motivates evaluation along dimensions such as **language coverage, fluency, prompt diversity, cross-lingual generalization, synthetic-data quality, and downstream transfer performance**.

2.6 Research Gap and Motivation

Collectively, these studies demonstrate that LLM capabilities are multidimensional and cannot be adequately characterized using conventional accuracy-based benchmarks alone. **Paper 2** emphasizes multimodal grounding and generation; **Papers 1 and 3** highlight causal reasoning and intervention-based evaluation; **Paper 4** demonstrates the importance of evaluating knowledge acquisition, forgetting, and internal concept representations; **Paper 5** motivates evaluation of foundation models through cross-task causal generalization and in-context learning; and **Pa-

per 6** highlights multilingual robustness and the relationship between synthetic-data quality and downstream model performance.

Despite these advances, these research directions are largely investigated independently. A comprehensive LLM evaluation framework should therefore move toward **multidimensional evaluation**, where models are assessed simultaneously across capability, reasoning, robustness, generalization, knowledge dynamics, multilinguality, and multimodal grounding. Moreover, benchmark performance alone may not distinguish genuine reasoning from memorization, superficial pattern matching, or generation of plausible but unsupported explanations.

This motivates the need for evaluation methodologies that incorporate **controlled interventions, counterfactual testing, distribution shifts, cross-task transfer, multilingual and multimodal settings, knowledge-retention analysis, and reliability assessment**. Such a framework would provide a more comprehensive understanding of not only *what* an LLM can accomplish, but also **how reliably, consistently, and generalizably it performs across different conditions**. The studies discussed above therefore provide the foundation for developing more rigorous and holistic approaches to LLM evaluation.

The proposed probabilistic Bayesian LLM-based evaluation framework addresses a fundamental limitation of lexical-overlap metrics such as BLEU and ROUGE: their sensitivity to surface-level word and n-gram matching. Rather than assuming that semantic similarity is proportional to lexical overlap, the framework uses an LLM to infer the semantic relationship between the prompt, generated response, and reference response, while Bayesian inference provides a probabilistic estimate of response quality and associated uncertainty. This enables semantically equivalent paraphrases to receive appropriate evaluations even when they exhibit limited lexical overlap, while uncertainty modeling can identify ambiguous or borderline cases that deterministic metrics cannot distinguish.

Bayesian uncertainty quantification itself is not a sufficient novelty claim. Our framework differs at the level of the probabilistic representation of evaluation outcomes. In particular, whereas recent approaches employ uniform or Dirichlet priors to model evaluation outcomes, our formulation introduces a Gaussian prior over a continuous latent semantic-quality variable. This distinction is motivated by the nature of our evaluation objective. BLEU and ROUGE primarily quantify lexical overlap, whereas semantic relatedness is inherently graded: two responses may differ substantially in their surface form while conveying nearly identical information, or may partially preserve the meaning of the reference. Representing evaluation quality as a continuous latent variable therefore allows the framework to model degrees of semantic agreement rather than restricting the evaluation to discrete outcome probabilities.

The Gaussian formulation additionally provides an explicit posterior distribution over semantic quality, allowing both the expected evaluation score and its uncertainty to be estimated. In the conjugate formulation considered in our framework, posterior parameters can be obtained analytically, avoiding iterative sampling for the Bayesian update. This provides a computationally tractable mechanism for uncertainty-aware semantic evaluation. We will revise the manuscript to make clear that the novelty does not lie in the use of Bayesian inference per se, but in the use of a continuous probabilistic semantic-quality formulation and its integration with LLM-based semantic evaluation. We will further provide an explicit theoretical and empirical comparison with the uniform/Dirichlet formulations used in recent Bayesian evaluation frameworks, including comparisons of outcome representation, posterior uncertainty, computational complexity, calibration, and robustness to semantically equivalent paraphrases.

3 Problem Formulation

Let $M = \{m_1, m_2, \ldots, m_K\}$ denote a set of K vision language models to be evaluated on a benchmark $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an input instance and y_i is its associated reference (gold) label or output. For a given model m_k and instance x_i, we define a binary or continuous *correctness indicator*

$$s_i^{(k)} = f\big(\hat{y}_i^{(k)}, y_i\big), \tag{1}$$

where $\hat{y}_i^{(k)}$ is the model's prediction and $f(\cdot, \cdot)$ is a scoring function (e.g., exact match, F1, a normalized rubric score, or a thresholded judge score). Conventional evaluation reports an aggregate

point estimate

$$\bar{s}^{(k)} = \frac{1}{N} \sum_{i=1}^{N} s_i^{(k)}, \tag{2}$$

which we refer to as the *frequentist performance estimate*. While $\bar{s}^{(k)}$ summarizes average behavior, it discards information about the variability of $s_i^{(k)}$ across instances and provides no native mechanism for expressing confidence in the estimate, particularly under small or non-i.i.d. benchmark samples. We instead treat $\bar{s}^{(k)}$ as an observed statistic generated by an underlying latent *performance parameter* $\theta^{(k)} \in [0, 1]$, which we infer probabilistically.

4 Methodology: Amortized Bayesian Reliability Estimation

We propose *Amortized Bayesian Reliability Estimation* (ABRE), a framework for quantifying the reliability of large language model (LLM) evaluation through posterior inference over a latent performance parameter. Rather than performing posterior inference independently for every model–benchmark pair, ABRE *amortizes* inference: a single neural network, trained once over a distribution of evaluation scenarios, maps any newly observed set of evaluation scores directly to an approximate posterior in a single forward pass. This is an instance of simulation-based (likelihood-free) neural posterior estimation (**?**), specialized to the setting of scalar model-reliability estimation.

The framework comprises five components: (i) a probabilistic model of LLM performance, (ii) a permutation-invariant representation of evaluation outcomes, (iii) an amortized inference network approximating the posterior, (iv) a reliability score with associated uncertainty derived from that posterior, and (v) posterior-based procedures for pairwise and multi-model comparison. **??** (if included) summarizes the pipeline; we describe each component below.

4.1 Problem Formulation

For model m_k, let $\theta^{(k)} \in [0, 1]$ denote a latent *reliability parameter*: the expected performance of m_k under the distribution from which evaluation instances are drawn. Given an observed evaluation set

$$\mathcal{D}^{(k)} = \{s_1^{(k)}, \ldots, s_N^{(k)}\}, \qquad s_i^{(k)} \in [0, 1], \tag{3}$$

our goal is not a single point estimate of performance but the full posterior

$$p\big(\theta^{(k)} \mid \mathcal{D}^{(k)}\big) \;\propto\; p\big(\mathcal{D}^{(k)} \mid \theta^{(k)}\big)\, p(\theta^{(k)}), \tag{4}$$

whose location summarizes expected performance and whose spread quantifies the epistemic uncertainty induced by a finite, possibly small or heterogeneous, evaluation set. We treat N as fixed per model in Section 4.1–Section 4.8 and relax this in Section 4.11.

4.2 Probabilistic Observation Model

We consider two observation models, selected according to the score type of the benchmark.

Binary correctness. For pass/fail scoring,

$$s_i^{(k)} \mid \theta^{(k)} \overset{\text{iid}}{\sim} \text{Bernoulli}\big(\theta^{(k)}\big), \qquad i = 1, \ldots, N. \tag{5}$$

Bounded continuous scores. For normalized rubric or judge-assigned scores $s_i^{(k)} \in (0, 1)$, we use a reparameterized Beta likelihood with mean $\theta^{(k)}$ and concentration $\kappa > 0$:

$$s_i^{(k)} \mid \theta^{(k)}, \kappa \overset{\text{iid}}{\sim} \text{Beta}\big(\theta^{(k)}\kappa,\, (1 - \theta^{(k)})\kappa\big). \tag{6}$$

Here $\theta^{(k)}$ retains a direct interpretation as expected score, while κ governs dispersion around it and is treated as a fixed hyperparameter or, when scores exhibit variable dispersion across instances, jointly inferred (Section 4.11 discusses subgroup-specific extensions). Both models share the conditional independence assumption

$$p\big(\mathcal{D}^{(k)} \mid \theta^{(k)}\big) = \prod_{i=1}^{N} p\big(s_i^{(k)} \mid \theta^{(k)}\big), \tag{7}$$

which lets ABRE treat discrete and continuous evaluation metrics within one formulation; instance-level correlation (e.g., shared context across multi-turn items) can be accommodated by the hierarchical extension of Section 4.11 rather than by this base model.

4.3 Prior Distribution

We place a conjugate Beta prior on the reliability parameter,

$$\theta^{(k)} \sim \mathrm{Beta}(\alpha_0, \beta_0), \qquad \mathbb{E}[\theta^{(k)}] = \frac{\alpha_0}{\alpha_0 + \beta_0}, \qquad \mathrm{Var}[\theta^{(k)}] = \frac{\alpha_0 \beta_0}{(\alpha_0 + \beta_0)^2(\alpha_0 + \beta_0 + 1)}. \quad (8)$$

The uninformative choice $\alpha_0 = \beta_0 = 1$ (uniform prior) is used unless historical evaluation data for related models or benchmarks motivate an informative prior via moment matching. Under the Bernoulli likelihood equation 5, conjugacy gives the closed-form posterior $\mathrm{Beta}(\alpha_0 + \sum_i s_i^{(k)},\ \beta_0 + N - \sum_i s_i^{(k)})$; under the Beta likelihood equation 6 with unknown or instance-varying κ, no closed form is generally available. This asymmetry is precisely what motivates amortized inference: rather than relying on conjugacy in the tractable case and repeated MCMC in the intractable case, ABRE learns a single inference map that applies uniformly to both.

4.4 Evaluation-Set Representation

Because the posterior in equation 4 depends on $\mathcal{D}^{(k)}$ only through the exchangeable collection of scores, any valid inference network must be invariant to the order in which instances are indexed. We therefore represent $\mathcal{D}^{(k)}$ through a permutation-invariant embedding

$$\mathbf{z}^{(k)} = g_\psi\big(\mathcal{D}^{(k)}\big) \in \mathbb{R}^d, \quad (9)$$

and consider two instantiations of g_ψ.

Fixed summary statistics. A hand-designed, permutation-invariant statistic,

$$\mathbf{z}^{(k)} = \Big[\, N,\ \bar{s}^{(k)},\ v^{(k)},\ q^{(k)}_{0.25},\ q^{(k)}_{0.50},\ q^{(k)}_{0.75} \,\Big], \quad (10)$$

where $\bar{s}^{(k)}$ and $v^{(k)}$ are the empirical mean and variance and $q^{(k)}_\tau$ the empirical τ-quantile of $\mathcal{D}^{(k)}$. This encoder has no learnable parameters ($\psi = \varnothing$) and serves as a low-variance baseline representation.

Learned DeepSets encoder. Following the DeepSets construction (**?**), which shows that every permutation-invariant function of a set admits a decomposition of the form below (under mild conditions),

$$\mathbf{z}^{(k)} = \rho_\psi\left(\frac{1}{N}\sum_{i=1}^{N} \phi_\psi\big(s_i^{(k)}\big)\right), \quad (11)$$

where $\phi_\psi : \mathbb{R} \to \mathbb{R}^{d'}$ embeds individual scores and $\rho_\psi : \mathbb{R}^{d'} \to \mathbb{R}^d$ produces the pooled representation. Because equation 11 averages over instances rather than concatenating them, its input dimensionality is independent of N, allowing a single trained network to process evaluation sets of arbitrary and unequal size. Unlike equation 10, this encoder can capture distributional features (e.g., multimodality or instance-difficulty structure) not summarized by low-order moments.

4.5 Amortized Posterior Inference

Given a representation $\mathbf{z}^{(k)}$, an inference head $h_\varphi : \mathbb{R}^d \to \mathbb{R}^2$ maps it to the natural parameters of an approximate Beta posterior,

$$\big(a^{(k)}, b^{(k)}\big) = h_\varphi\big(\mathbf{z}^{(k)}\big), \qquad \hat{\alpha}^{(k)} = \mathrm{softplus}\big(a^{(k)}\big) + \epsilon, \quad \hat{\beta}^{(k)} = \mathrm{softplus}\big(b^{(k)}\big) + \epsilon, \quad (12)$$

with $\epsilon > 0$ a small constant ensuring $\hat{\alpha}^{(k)}, \hat{\beta}^{(k)} > 0$. The approximate posterior is then

$$q_{\psi,\varphi}\big(\theta^{(k)} \mid \mathcal{D}^{(k)}\big) = \mathrm{Beta}\big(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}\big) \ \approx\ p\big(\theta^{(k)} \mid \mathcal{D}^{(k)}\big), \quad (13)$$

where we write $q_{\psi,\varphi}$ (rather than q_φ) to make explicit that the encoder parameters ψ and the head parameters φ are optimized jointly; we abbreviate this as q_Φ with $\Phi = (\psi, \varphi)$ below. Choosing a Beta family for q_Φ matches the support of $\theta^{(k)}$ and the conjugate form of equation 4 under equation 5, so that in the conjugate case the amortized posterior can, in principle, recover the exact posterior exactly; more expressive families (e.g., normalizing flows on $[0, 1]$) are a straightforward extension when the true posterior is expected to be strongly multimodal or skewed beyond what a Beta can represent. Once trained, q_Φ requires a single forward pass through g_ψ and h_φ to produce a posterior for a previously unseen evaluation set, in contrast to the per-dataset optimization or sampling required by conventional Bayesian inference.

4.6 Training the Amortized Inference Network

The parameters $\Phi = (\psi, \varphi)$ are learned from a corpus of training tasks $\mathcal{T} = \{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$, each paired with a known latent parameter. Training tasks may be (a) fully synthetic, simulated directly from the generative model of Section 4.2–Section 4.3, or (b) constructed from historical evaluation logs whose latent reliability is estimated separately via high-fidelity (non-amortized) Bayesian inference. In the synthetic case, for $j = 1, \ldots, M$:

$$\theta_j \sim \mathrm{Beta}(\alpha_0, \beta_0), \qquad N_j \sim p(N), \qquad s_{ij} \overset{\text{iid}}{\sim} p(s \mid \theta_j), \quad i = 1, \ldots, N_j, \tag{14}$$

where $p(N)$ is a distribution over evaluation-set sizes (matched, in practice, to the range $N \in \{10, 25, 50, 100, 250, 500, 1000\}$ used at test time) so that the trained network generalizes across sample sizes rather than overfitting to a single N. This yields simulated pairs $\{(\mathcal{D}_j, \theta_j)\}_{j=1}^M$.

Posterior density objective. The primary training objective is the negative log density that the amortized posterior assigns to the true generating value θ_j,

$$\mathcal{L}_{\mathrm{NLL}}(\Phi) = -\frac{1}{M} \sum_{j=1}^{M} \log q_\Phi\big(\theta_j \mid \mathcal{D}_j\big). \tag{15}$$

Minimizing equation 15 is equivalent, in expectation over the generative model, to minimizing the forward KL divergence $\mathbb{E}_{\mathcal{D}}\big[D_{\mathrm{KL}}\big(p(\theta \mid \mathcal{D}) \,\|\, q_\Phi(\theta \mid \mathcal{D})\big)\big]$ up to a constant independent of Φ, which is the standard consistency argument for neural posterior estimation (**?**): at the global optimum over an infinitely flexible q_Φ, $\mathcal{L}_{\mathrm{NLL}}$ is minimized exactly when $q_\Phi(\theta \mid \mathcal{D}) = p(\theta \mid \mathcal{D})$ for (almost) every $\mathcal{D}$ in the support of the training distribution.

Calibration regularization. Because the Beta family and finite network capacity limit exact recovery, we additionally regularize using a reference posterior $p(\theta \mid \mathcal{D}_j)$ obtained by exact conjugate updating (when available) or MCMC (otherwise), evaluated on a held-out calibration subset $\mathcal{T}_{\mathrm{cal}} \subset \mathcal{T}$:

$$\mathcal{L}_{\mathrm{KL}}(\Phi) = \frac{1}{|\mathcal{T}_{\mathrm{cal}}|} \sum_{j \in \mathcal{T}_{\mathrm{cal}}} D_{\mathrm{KL}}\big(p(\theta \mid \mathcal{D}_j) \,\|\, q_\Phi(\theta \mid \mathcal{D}_j)\big), \tag{16}$$

which, for two Beta distributions, has closed form in terms of their parameters. The combined objective is

$$\mathcal{L}(\Phi) = \mathcal{L}_{\mathrm{NLL}}(\Phi) + \lambda\, \mathcal{L}_{\mathrm{KL}}(\Phi), \qquad \lambda \geq 0, \tag{17}$$

with λ selected on a validation split. Section 4.13 studies the effect of $\lambda = 0$ versus $\lambda > 0$.

4.7 Bayesian Reliability Score

Given a trained network and an observed evaluation set $\mathcal{D}^{(k)}$, we define the **Amortized Bayesian Reliability Score (ABRS)** as the posterior mean under q_Φ,

$$\mathrm{ABRS}^{(k)} = \mathbb{E}_{q_\Phi(\theta^{(k)} \mid \mathcal{D}^{(k)})}\big[\theta^{(k)}\big] = \frac{\hat{\alpha}^{(k)}}{\hat{\alpha}^{(k)} + \hat{\beta}^{(k)}}, \tag{18}$$

with associated posterior variance

$$\mathrm{Var}_{q_\Phi}\big[\theta^{(k)}\big] = \frac{\hat{\alpha}^{(k)} \hat{\beta}^{(k)}}{\big(\hat{\alpha}^{(k)} + \hat{\beta}^{(k)}\big)^2 \big(\hat{\alpha}^{(k)} + \hat{\beta}^{(k)} + 1\big)} \tag{19}$$

and equal-tailed $(1-\gamma)$ credible interval

$$\mathrm{CI}^{(k)}_{1-\gamma} = \Big[F^{-1}_{\hat{\alpha}^{(k)},\hat{\beta}^{(k)}}(\gamma/2),\ F^{-1}_{\hat{\alpha}^{(k)},\hat{\beta}^{(k)}}(1-\gamma/2) \Big], \tag{20}$$

where $F^{-1}_{\hat{\alpha}^{(k)},\hat{\beta}^{(k)}}$ is the quantile function of $\mathrm{Beta}(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)})$. We report $\mathrm{ABRS}^{(k)}$ together with equation 19 or equation 20 rather than in isolation; a point estimate alone discards the information the posterior was constructed to provide.

4.8 Reliability as the Joint Report of Estimate and Uncertainty

We treat "reliability" as a property of the *joint* report $(\mathrm{ABRS}^{(k)}, \mathrm{Var}_{q_\Phi}[\theta^{(k)}])$ rather than of the point estimate alone. A narrow posterior indicates that $\mathcal{D}^{(k)}$ constrains $\theta^{(k)}$ strongly under the assumed observation model; it is not, by itself, evidence of high performance, and we deliberately avoid conflating the two by always reporting them as separate quantities.

We assess whether the amortized posterior's stated uncertainty is trustworthy via empirical coverage on independently simulated test tasks $\{(\mathcal{D}_j, \theta_j)\}_{j=1}^{M'}$ with known ground truth:

$$\mathrm{Coverage}_{1-\gamma} = \frac{1}{M'} \sum_{j=1}^{M'} \mathbb{I}\big[\theta_j \in \mathrm{CI}_{1-\gamma,j}\big]. \tag{21}$$

A well-calibrated q_Φ satisfies $\mathrm{Coverage}_{1-\gamma} \approx 1-\gamma$ simultaneously across nominal levels $1-\gamma$ (e.g., $\{0.5, 0.8, 0.9, 0.95\}$); we report coverage curves rather than a single level to detect systematic over- or under-confidence.

4.9 Bayesian Pairwise Model Comparison

For two models m_a, m_b with amortized posteriors $q_\Phi(\theta^{(a)} \mid \mathcal{D}^{(a)})$ and $q_\Phi(\theta^{(b)} \mid \mathcal{D}^{(b)})$, we approximate the posterior over their performance difference $\Delta^{(a,b)} = \theta^{(a)} - \theta^{(b)}$ by Monte Carlo: drawing independent samples $\theta^{(a)}_t \sim q_\Phi(\theta^{(a)} \mid \mathcal{D}^{(a)})$ and $\theta^{(b)}_t \sim q_\Phi(\theta^{(b)} \mid \mathcal{D}^{(b)})$ for $t = 1, \ldots, T$, and forming $\Delta^{(a,b)}_t = \theta^{(a)}_t - \theta^{(b)}_t$. The independence assumption reflects that $\theta^{(a)}$ and $\theta^{(b)}$ are inferred from disjoint evaluation runs; if the two models were scored on identical instances with shared judge noise, this independence would need to be relaxed (e.g., via paired evaluation and a joint amortized posterior).

From these samples we estimate the posterior probability of superiority,

$$P\big(\theta^{(a)} > \theta^{(b)} \mid \mathcal{D}^{(a)}, \mathcal{D}^{(b)}\big) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\big[\Delta^{(a,b)}_t > 0\big], \tag{22}$$

and, given a practical-equivalence margin $\epsilon > 0$ fixed *before* observing $\mathcal{D}^{(a)}, \mathcal{D}^{(b)}$, the region-of-practical-equivalence (ROPE) probability

$$P\big(|\Delta^{(a,b)}| < \epsilon\big) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\big[|\Delta^{(a,b)}_t| < \epsilon\big]. \tag{23}$$

Reporting both equation 22 and equation 23 distinguishes "m_a is probably better" from "m_a and m_b are probably indistinguishable at a practically meaningful scale," which a significance test alone does not.

4.10 Multi-Model Posterior Comparison

For K models evaluated jointly, we obtain independent amortized posteriors $q_\Phi(\theta^{(k)} \mid \mathcal{D}^{(k)})$, $k = 1, \ldots, K$, and draw joint samples $\big(\theta^{(1)}_t, \ldots, \theta^{(K)}_t\big)$, $t = 1, \ldots, T$, by sampling each component independently. For each draw, we compute the descending-order rank of each model,

$$R^{(k)}_t = 1 + \sum_{k' \neq k} \mathbb{I}\big[\theta^{(k')}_t > \theta^{(k)}_t\big], \tag{24}$$

with ties (a measure-zero event under continuous q_Φ, but possible under numerical rounding) broken uniformly at random. The empirical distribution of $R_t^{(k)}$ over $t = 1, \ldots, T$ summarizes rank uncertainty for model m_k and can be reported as, e.g., $P(R^{(k)} = 1)$ ("probability of being best") or the full rank histogram, in place of a single deterministic leaderboard position.

4.11 Hierarchical Amortized Bayesian Evaluation

Benchmarks are frequently partitioned into J subgroups (tasks, languages, domains, or subsets) of possibly unequal size. Let $\theta_j^{(k)}$ denote model m_k's reliability on subgroup j, and let $\mathcal{D}_j^{(k)}$ denote the corresponding scores. We assume subgroup-level parameters are drawn from a model-specific population distribution,

$$\theta_j^{(k)} \mid \alpha^{(k)}, \beta^{(k)} \overset{\text{iid}}{\sim} \mathrm{Beta}\big(\alpha^{(k)}, \beta^{(k)}\big), \qquad j = 1, \ldots, J, \tag{25}$$

with $(\alpha^{(k)}, \beta^{(k)})$ themselves given hyperpriors, inducing partial pooling across subgroups: subgroups with few instances borrow statistical strength from the population-level distribution, while subgroups with abundant data are dominated by their own likelihood. The amortized network is extended to take the collection of per-subgroup representations as input,

$$q_\Phi\big(\theta_{1:J}^{(k)}, \alpha^{(k)}, \beta^{(k)} \mid \mathcal{D}_{1:J}^{(k)}\big), \tag{26}$$

approximating the joint posterior over subgroup-level and population-level parameters simultaneously. In practice, g_ψ from Section 4.4 is applied to each subgroup independently (producing $\mathbf{z}_j^{(k)}$), and a second, subgroup-permutation- invariant pooling layer aggregates $\{\mathbf{z}_j^{(k)}\}_{j=1}^J$ into a population-level summary that conditions the prediction of $(\alpha^{(k)}, \beta^{(k)})$, while each $\mathbf{z}_j^{(k)}$ together with this population summary conditions the prediction of $\theta_j^{(k)}$'s posterior. This is particularly beneficial when individual subgroups are too small for the base (non-hierarchical) model of Section 4.2 to yield informative posteriors on their own.

4.12 Experimental Protocol

We evaluate ABRE along four axes:

1. **Posterior accuracy**: agreement between q_Φ and a reference posterior from exact conjugate updating (Bernoulli likelihood) or high-fidelity MCMC (Beta likelihood), measured by KL divergence and by discrepancy in posterior mean and variance.
2. **Calibration**: empirical coverage (Equation (21)) at nominal levels $\{0.5, 0.8, 0.9, 0.95\}$.
3. **Reliability estimation**: ABRS accuracy and posterior width as functions of evaluation-set size N, score variance, and subgroup heterogeneity.
4. **Computational efficiency**: wall-clock inference latency and total compute of ABRE versus per-dataset conjugate updating and per-dataset MCMC.

We evaluate across $N \in \{10, 25, 50, 100, 250, 500, 1000\}$ and multiple score-heterogeneity regimes, using both synthetic tasks with known ground-truth θ and real evaluation logs from multiple LLMs across multiple benchmarks. For the latter, held-out model–benchmark combinations are withheld from training entirely, so that generalization is assessed on scenarios the network has not amortized over during training.

4.13 Ablation Studies

To isolate the source of any improvement, we ablate:

1. **No amortization**: independent per-dataset Bayesian inference (exact or MCMC), serving as the non-amortized reference.
2. **Mean-only representation**: $\mathbf{z}^{(k)} = [\bar{s}^{(k)}]$, testing whether higher-order statistics matter.
3. **Summary-statistic encoder**: g_ψ as in equation 10.

4. **Learned set encoder**: g_ψ as in equation 11.
5. **Prior sensitivity**: weakly informative ($\alpha_0 = \beta_0 = 1$) versus informative priors constructed from historical data.
6. **Calibration objective**: $\lambda = 0$ versus $\lambda > 0$ in equation 17.

Comparing (1) against (3)/(4) isolates the effect of amortization itself; comparing (2) against (3)/(4) isolates the effect of the representation; and comparing (5)-(6) isolates sensitivity to modeling choices orthogonal to the network architecture.

4.14 Experimental Setup

Models. We instantiate the proposed framework on a set of representative language models spanning different architectures and pretraining objectives, including **PolyGemma**, **BERT**, and **Polyglot**. [Add exact model variants/sizes and checkpoints, e.g., `google/paligemma-3b`, `bert-base-uncased`, `EleutherAI/polyglot-ko-1.3b`, and specify quantization/precision settings used at inference time.]

Benchmarks. Evaluation is conducted on [list benchmark datasets $\mathcal{D}_1, \ldots, \mathcal{D}_M$, task types (classification, generation, multilingual QA, etc.), and dataset sizes N per benchmark]. For each benchmark, per-instance scores $s_i^{(k)}$ are computed using the scoring function $f(\cdot, \cdot)$ appropriate to the task (e.g., exact-match accuracy, F1, or an LLM-judge rubric score normalized to $[0, 1]$).

Baselines. We compare the proposed Bayesian Reliability Score against conventional point-estimate baselines, including: (i) raw accuracy/F1 (Equation (2)); (ii) bootstrap confidence intervals computed via resampling of $\mathbf{s}^{(k)}$; and (iii) normal-approximation (Wald) confidence intervals. This allows us to directly assess whether the Bayesian posterior uncertainty (**??**) captures reliability information that is obscured or misrepresented by these conventional approaches, particularly in small-sample or high-variance regimes.

Inference Details. Posterior inference is performed using [PyMC / NumPyro / Stan — specify], with $C = 4$ chains, $T = 2000$ post-warmup samples per chain, and 1000 warmup (tuning) iterations. Convergence is verified via $\hat{R} < 1.01$ and effective sample size (ESS) diagnostics for all reported parameters. All experiments are repeated across $[R]$ random seeds / benchmark shuffles to assess the stability of both the point estimates and the induced posterior uncertainty.

Evaluation Metrics. We report: (i) the Bayesian Reliability Score and credible intervals (**??**, **??**) per model and benchmark; (ii) posterior standard deviation as a reliability indicator (**??**); (iii) pairwise probability-of-superiority and ROPE-based practical equivalence statistics (Section 4.9) for all model pairs; and (iv) calibration diagnostics comparing nominal credible-interval coverage to empirical coverage under repeated benchmark subsampling.

4.15 Algorithm Summary

Algorithm 1 summarizes the end-to-end procedure for computing the Bayesian Reliability Score for a single model on a given benchmark.

References

A Appendix

You may include other additional sections here.

Algorithm 1 Bayesian Reliability Score Estimation

Require: Benchmark $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, model m_k, scoring function $f(\cdot, \cdot)$, prior hyperparameters (α_0, β_0)

Ensure: Bayesian Reliability Score $\text{BRS}^{(k)}$, credible interval $\text{CI}_{1-\gamma}^{(k)}$

1: Generate predictions $\hat{y}_i^{(k)} \leftarrow m_k(x_i)$ for all $i = 1, \ldots, N$
2: Compute per-instance scores $s_i^{(k)} \leftarrow f(\hat{y}_i^{(k)}, y_i)$
3: **if** scores are binary **then**
4: Compute posterior in closed form via **??**
5: **else**
6: Draw posterior samples $\{\theta_{(t)}^{(k)}\}_{t=1}^{T}$ via MCMC (NUTS) using Equation (6) and Equation (8)
7: Verify convergence ($\hat{R} < 1.01$, adequate ESS)
8: **end if**
9: Compute $\text{BRS}^{(k)} \leftarrow \mathbb{E}[\theta^{(k)} \mid \mathbf{s}^{(k)}]$ (**??**)
10: Compute $\text{CI}_{1-\gamma}^{(k)}$ from posterior quantiles (**??**)
11: **return** $\text{BRS}^{(k)}, \text{CI}_{1-\gamma}^{(k)}$