Embracing Ambiguity: Bayesian Nonparametrics and Stakeholder Participation for Ambiguity-Aware Safety Evaluation

Yanan Long
ylong@uchicago.edu
StickFlux Labs

Abstract

Safety evaluations of generative models often collapse nuanced behaviour into a single number computed for a single decoding configuration. Such *point estimates* obscure tail risks, demographic disparities, and the existence of multiple near-optimal operating points—phenomena collectively known as the Rashomon or predictive multiplicity effect. We propose a unified framework that embraces multiplicity by modelling the distribution of harmful behaviour across the entire space of decoding knobs and prompts, quantifying risk through tail-focused metrics, and integrating stakeholder preferences. Our technical contributions are threefold: (i) we formalise *decoding Rashomon sets*—regions of knob space whose risk is near-optimal under given criteria—and measure their size and disagreement; (ii) we develop a dependent Dirichlet process mixture with stakeholder-conditioned, prompt-aware stick-breaking weights to learn multi-modal harm surfaces; and (iii) we introduce an active sampling and calibration pipeline that uses Bayesian deep learning surrogates and conformal wrappers to explore knob space efficiently while maintaining finite-sample coverage guarantees. The framework supports simulated stakeholder participation: synthetic stakeholders draw prompts from a topic mixture anchored to real datasets and rate outputs according to demographic-specific sensitivities. We demonstrate on synthetic and real LLM evaluations that our method reveals hidden failure modes, quantifies disagreement across stakeholders, and identifies safe operating regions that single-point evaluations miss. Our approach bridges multiplicity theory, Bayesian nonparametrics, uncertainty quantification, and participatory AI, paving the way for trustworthy deployment of generative models.

Introduction

Large generative models now participate in education, healthcare, policy support and creative industries, yet they often produce outputs that are toxic, biased or factually incorrect (Bender et al. 2021; Weidinger et al. 2022). Evaluation protocols typically fix a handful of decoding settings—temperature, top-p, repetition penalty or model family indicators—generate a few samples per prompt, compute average toxicity scores, and declare success if the average risk is below a threshold. This practice tacitly assumes that (i) there is a unique *best* operating point and (ii) the mean is the only relevant risk functional. Both assumptions are problematic. Real tasks exhibit *predictive multiplicity*: many decoding configurations yield similar average metrics but differ dramatically in tail behaviour and fairness across demographic slices (Breiman 2001). Policies that ignore such multiplicity can inadvertently select an operating point with unacceptable tail risk or bias. Recent work in interpretable machine learning has formalised multiplicity through *Rashomon sets*: the collection of models with near-optimal performance (Rudin et al. 2019). Motivated by this perspective, we argue that safety evaluation of generative models should recognise and report multiplicity rather than suppress it.

Our goal is thus not to return a single toxicity score but to infer a *posterior distribution over risk surfaces* across knob space and prompts, then compute tail-sensitive and stakeholder-specific risk summaries. This requires three innovations. First, we formalise *decoding Rashomon sets* as subsets of knob space where a risk functional (mean, conditional value at risk, or disparity gap) is within ε of its optimum; we propose measures of their size and how they differ across stakeholders. Second, we develop a dependent Dirichlet process (DDP) mixture model to capture multi-modal harm distributions that vary with both decoding settings and prompt features. Unlike parametric logistic surfaces, the DDP yields a flexible family of conditional distributions and naturally expresses multiplicity via its posterior. Third, we design an active sampling and stakeholder simulation pipeline that calibrates automated toxicity detectors via synthetic human ratings, uses Bayesian deep neural networks to guide exploration of knob space, and employs conformal prediction wrappers for finite-sample coverage.

Our framework treats stakeholders as first-class participants who define prompt and knob distributions and evaluation criteria. Because recruiting diverse raters is expensive and ethically sensitive, we present a simulation protocol grounded in public datasets (RealToxicityPrompts, Civil Comments, BOLD, SBIC) where synthetic stakeholders draw prompts from topic mixtures and rate outputs with demographic-specific noise, bias and severity. We calibrate automated judges using these labels to obtain human-aligned harm probabilities. We then fit the DDP mixture to learn multi-modal harm surfaces and compute risk measures such

1st Workshop on Navigating Model Uncertainty and the Rashomon Effect: From Theory and Tools to Applications and Impact

as conditional value at risk (CVaR), worst-slice gap, safe-set volume, and disagreement index. By integrating stakeholder policies over knob space, we report stakeholder-specific risk and identify knob regions that are simultaneously safe and high-utility.

Recent advances in conformal prediction and Bayesian evaluation of LLMs motivate and complement our approach. Adaptive conformal methods build local partitions of the predictor space and yield group-conditional coverage (Kim, O'Hagan, and Ročková 2025), while selective conformal uncertainty (SConU) uses significance tests to detect violations of exchangeability and improve miscoverage rates (Wang et al. 2025). Our framework leverages these ideas by wrapping DDP predictions in conformal bands to ensure valid uncertainty quantification even when prompts and knobs change. Concurrently, Bayesian evaluation of LLM behaviour models stochastic generation as a Beta–Binomial process and uses sequential sampling to reduce evaluation cost (Longjohn et al. 2025). We extend this line of work to multi-modal distributions and knob-continuous risk surfaces. Finally, calls to human-centered uncertainty quantification emphasise the need to account for aleatoric and distributional uncertainty and to measure utility for actual users (Devic et al. 2025). Our stakeholder simulation responds directly by placing value preferences and usage policies at the core of the risk computation.

The remainder of the paper is organised as follows. We first provide a formal problem formulation, defining prompts, knob space, harm variables and stakeholders. We then lay out our four-stage pipeline: space-filling design and active sampling, stakeholder prior elicitation and policy specification, simulated judging and calibration, and Bayesian inference with risk reporting. We present the full DDP generative model, describe Bayesian deep learning surrogates for active sampling and logistic calibration, and derive conformal wrappers for finite-sample coverage. We discuss simulation details and relate our work to multiplicity, fairness and uncertainty literature. We conclude with experiments on synthetic and real LLM evaluations and outline future directions.

Formal Problem Statement and Risk Functionals

Generative model and decoding knobs

Let $x \in \mathcal{X} \subset \mathbb{R}^d$ denote a vector of decoding settings, such as temperature, nucleus sampling probability, repetition penalty, model family indicators, or refusal filter toggles. The generative model π defines a stochastic mapping from prompts $p \in \mathcal{P}$ and knobs x to outputs $y \sim \pi(\cdot \mid p, x)$. Because decoding is random, each (p, x) pair induces a full distribution over outputs. We assume a finite set of prompts $\{p_i\}_{i=1}^{I}$ drawn either from a stakeholder-defined distribution $p(p \mid s)$ or from a simulation process described later.

We are interested in harmful behaviour captured by a binary or continuous harm score $H(y)$. For example, H could be a toxicity probability, a measure of privacy leakage, or the indicator of a jailbreak. In practice H is not directly observable but is estimated via a judge—either an automated classifier or a human rater. We denote the calibrated harm score by $\tilde{h}(y) \in [0, 1]$ and treat it as an estimate of $\Pr(\text{harm} \mid y)$.

Stakeholders

A stakeholder s is characterised by three ingredients: (i) a distribution over prompts $p(p \mid s)$ reflecting their domain of interest; (ii) a distribution over decoding knobs $p(x \mid s)$ reflecting typical usage or policy constraints; and (iii) a risk threshold or preference functional. Stakeholders may represent content moderators, product managers, developers, or different demographic user groups. When real participants are unavailable, we simulate stakeholders by sampling their prompt topics, knob policies and sensitivities from priors anchored to public datasets and demographic research (see Section).

Risk functionals and multiplicity metrics

Let $Z_{p,x}$ denote the predictive harm random variable at prompt p and knob setting x. The *mean harm surface* is $\mu(p, x) = \mathbb{E}[Z_{p,x}]$. To capture tail risk, we define the conditional value at risk (CVaR) at level $\alpha \in (0, 1)$,

$$\mathrm{CVaR}_\alpha(p, x) = \frac{1}{1-\alpha}\,\mathbb{E}\big[Z_{p,x} \,\big|\, Z_{p,x} \geq F^{-1}_{p,x}(\alpha)\big], \quad (1)$$

where $F_{p,x}$ is the distribution function of $Z_{p,x}$ (Rockafellar and Uryasev 2000). CVaR measures expected harm in the worst $1-\alpha$ fraction of outcomes. For fairness, let $\mathcal{G}$ be a set of demographic slices (e.g., identity groups). The harm rate for group g at x is $\mu_g(x) = \mathbb{E}[Z_{p,x} \mathbb{1}\{p \in g\}] / \Pr(p \in g)$. The *worst-slice gap* is

$$\mathrm{Gap}(x) = \max_{g \in \mathcal{G}} \mu_g(x) - \min_{g \in \mathcal{G}} \mu_g(x). \quad (2)$$

Finally, for a stakeholder s with knob policy $p(x \mid s)$, the expected harm is

$$R_s = \int_{\mathcal{X}} \mu(p(s), x)\, p(x \mid s)\, \mathrm{d}x, \quad (3)$$

where $p(s)$ indicates that the prompt distribution depends on s. These functionals define the targets we wish to estimate.

Decoding Rashomon sets and disagreement metrics. Fix a risk functional $R(x)$ (e.g., $\mathrm{CVaR}_\alpha(x)$). The ε-Rashomon set is

$$\mathcal{R}_\varepsilon = \{x \in \mathcal{X} : R(x) \leq R^\star + \varepsilon\}, \quad (4)$$

where $R^\star = \min_x R(x)$ or any baseline. We measure multiplicity via the *safe volume*

$$\mathrm{Vol}_\varepsilon = \mathrm{vol}\big(\mathcal{R}_\varepsilon\big)/\mathrm{vol}(\mathcal{X}), \quad (5)$$

where vol denotes Lebesgue measure, and the *disagreement index*

$$\mathrm{Disagree}(x) = \mathrm{Var}_s\big(R_s(x)\big), \quad (6)$$

which quantifies how stakeholder risk differs at x. Posterior distributions over $\mathcal{R}_\varepsilon$ and Vol_ε capture epistemic uncertainty about multiplicity.

These risk functionals emphasise distinct aspects of harmful behaviour. The mean $\mu(p, x)$ measures the expected harm across stochastic generations, integrating over aleatoric uncertainty; however, for rare but catastrophic failures the mean

Stage 1: Space-filling design and active sampling
Stage 2: Prior elicitation and stakeholder policies
Stage 3: Judging, calibration, and simulated rating
Stage 4: Bayesian inference with DDP mixtures, Bayesian deep surrogates and conformal wrappers

Figure 1: Schematic overview of the four-stage evaluation pipeline. The pipeline begins with a space-filling design over the knob space and active sampling to select informative configurations. Stakeholders elicit priors and specify policies over prompts and knobs. Judges (automated and simulated stakeholders) produce calibrated harm labels. A dependent Dirichlet process mixture is fitted to the data, a Bayesian deep surrogate guides further sampling, and conformal methods wrap the posterior to provide finite-sample guarantees.

can be small even when the tail is unacceptable. CVaR_α therefore focuses on the worst $1-\alpha$ fraction of outcomes and is widely used in finance and risk management as a coherent risk measure (Rockafellar and Uryasev 2000). In our context, $\mathrm{CVaR}_{0.95}(p, x)$ quantifies expected toxicity among the top 5% most harmful completions for a prompt/knob pair. The worst-slice gap (2) quantifies fairness by comparing the highest and lowest mean harm rates across demographic slices $\mathcal{G}$; a large gap indicates that some identity groups are disproportionately harmed. Stakeholder risk R_s aggregates the mean harm over a stakeholder's knob policy and prompt distribution, weighting $\mu(p, x)$ by how likely the stakeholder is to encounter each configuration.

Besides these metrics, we also consider exceedance probabilities and quantiles. Let $q_\beta(p, x)$ denote the β-quantile of $Z_{p,x}$ (the inverse of $F_{p,x}$); then the exceedance probability $\Pr(Z_{p,x} \geq \tau)$ and quantile $q_\beta(p, x)$ can be estimated from posterior draws of Z. These quantities support threshold-based safety policies (e.g., "at most 5% of completions may exceed a toxicity threshold") and can be used to define Rashomon sets based on exceedance constraints.

When summarising risk across prompts, we integrate over the stakeholder's prompt distribution: $\bar{\mu}(x) = \mathbb{E}_{p\sim p(p|s)}[\mu(p, x)]$ and analogously $\overline{\mathrm{CVaR}}_\alpha(x)$. These collapsed surfaces drive the stakeholder risk (3) and are crucial when stakeholders have different prompt preferences.

The safe volume (5) captures how much of knob space satisfies a safety criterion. A large safe volume indicates that many settings are essentially equivalent in risk (high multiplicity), while a tiny safe volume suggests a narrow "sweet spot." The disagreement index (6) measures how stakeholder utilities diverge; high variance implies that different groups perceive the same knob setting very differently, signalling value multiplicity. In Section we describe simulation protocols that induce such divergences.

Methodology

Our evaluation pipeline comprises four stages (Figure 1).

Stage 1: Design, active sampling and prompts

Given knob domain $\mathcal{X}$ and budget B, we first select design points $\{x_j\}_{j=1}^N$ and replicates R_j for each x_j. A space-filling Sobol or Latin hypercube design covers the domain initially. We then adaptively refine the design by exploiting posterior uncertainty from early experiments: a Bayesian deep neural network (BDN) surrogate predicts harm $\hat{\mu}(x)$ with uncertainty $\hat{\sigma}(x)$ and selects new x to maximise acquisition functions such as uncertainty-weighted distance to the risk threshold or "straddle" for level-set estimation. This active sampling reduces the number of required model calls.

Each prompt p_i is drawn according to the stakeholder prompt distribution or, for simulations, from a mixture of harm topics anchored to public corpora. For each (p_i, x_j) pair we generate $R_{i,j}$ independent outputs y_{ijr} using the generative model π, capturing the stochasticity of decoding.

Stage 2: Prior elicitation and stakeholder policies

Stakeholders elicit prior beliefs about harm rates and specify usage policies over knobs and prompts. For simulated stakeholders, we draw knob policies $p(x \mid s)$ from Beta or Dirichlet distributions reflecting preferences (e.g., high temperature or low top-p), and sample prompt topics from group-specific Dirichlet mixtures over harm categories (hate speech, harassment, misinformation). Stakeholder severities and biases are encoded through priors on a latent severity shift parameter b_s and sensitivity a_s in the observation model. When stakeholders provide intuitive summaries like "outputs are harmful about 10% of the time," we solve for Beta prior parameters (a, b) matching the mean and credible interval; such transformations appear in Stage 2 of the original draft and remain unchanged here.

Stage 3: Judging, calibration and simulation

The raw harm score $H(y)$ is estimated via judges. Automated toxicity detectors provide initial scores $J(y) \in [0, 1]$. To account for miscalibration and bias, we collect a subset of human or simulated ratings $r(y) \in \{1, \ldots, L\}$ or pairwise preferences. For simulated stakeholders, we model a latent harm severity $\zeta(y)$ and generate ratings via an ordinal logistic model:

$$\Pr\big(r(y) \leq \ell \mid \zeta(y), s\big) = \sigma(\tau_{s,\ell} - a_s\,\zeta(y) - b_s)\,, \qquad (7)$$

where σ is logistic, a_s the stakeholder sensitivity, b_s severity bias and $\tau_{s,\cdot}$ thresholds. Pairwise preferences are sampled via a Bradley–Terry model,

$$\Pr\big(y_a \succ y_b \mid s\big) = \sigma\Big(a_s[\zeta(y_a) - \zeta(y_b)] + b_s\Big). \qquad (8)$$

We calibrate the automated judge using isotonic regression or Bayesian logistic calibration: we fit a mapping $\tilde{h}(y) = f(J(y); \beta)$ with posterior over β using the labelled subset and propagate uncertainty to the BNP model. If multiple automated judges exist, we combine them via a Bayesian last-layer ensemble.

Stage 4: BNP inference with stakeholder-conditioned DDP mixtures

The core of our approach is a flexible generative model for harm conditional on knobs x, prompts p and stakeholders s. We posit a latent mixture structure: for each output y we draw a component assignment $z \in \{1, \dots, K\}$, a harm parameter θ_k determining the distribution of $\tilde{h}(y)$, and mixture weights $\pi_k(x, p, s)$ that depend on the knob, prompt and stakeholder. Our generative model is:

$$z_{ijr} \sim \text{Categorical}(\pi_1(x_j, p_i, s), \dots, \pi_K(x_j, p_i, s)), \tag{9}$$

$$\tilde{h}(y_{ijr}) \mid z_{ijr} = k \sim \begin{cases} \text{Bernoulli}(\theta_k), & \text{binary harm,} \\ \text{Beta}(\eta_k, \lambda_k), & \text{continuous harm.} \end{cases} \tag{10}$$

The base measure H for θ_k can be $\text{Beta}(a, b)$; for Beta-distributed harm we can set $(\eta_k, \lambda_k) \sim \text{Gamma}$. The mixture weights follow a logistic stick-breaking construction:

$$v_k(x, p, s) = \sigma\big(g_k(x, p, s)\big), \quad \pi_k = v_k \prod_{h<k} (1 - v_h), \tag{11}$$

$$g_k(x, p, s) = \alpha_k^\top \phi(x) + \beta_k^\top \psi(p) + \delta_k^\top \rho(s), \tag{12}$$

where $\phi(x)$ is a basis expansion of knobs (splines, random Fourier features), $\psi(p)$ are prompt features (topic embeddings or bag-of-words), and $\rho(s)$ encodes stakeholder identity. The gating coefficients $(\alpha_k, \beta_k, \delta_k)$ follow Gaussian priors with variance hyperparameter τ^2. When the number of components K is large, this truncated mixture approximates a full DDP (Ren, Dunson et al. 2011). If knob or prompt features interact multiplicatively, we can include tensor product terms or model g_k via a Bayesian neural network.

Inference. We fit this model using either Hamiltonian Monte Carlo (HMC) or stochastic variational inference (SVI). HMC provides accurate posterior samples but scales to at most thousands of observations; SVI scales to tens of thousands by amortising updates. We integrate over latent assignments z in variational updates or sample them via Gibbs. We monitor effective sample sizes and Gelman–Rubin statistics for convergence. Posterior predictive surfaces $\hat{\mu}(p, x)$, $\widehat{\text{CVaR}}_\alpha(p, x)$ and $\widehat{\text{Gap}}(x)$ are computed from draws of the mixture parameters.

Bayesian deep surrogates. Evaluating the LLM π across a large design is costly because each $\{(p, x)\}$ pair requires sampling multiple outputs and calibrating their harms. To amortise this cost, we may resort to a *Bayesian deep neural network* (BDN) surrogate $f_\omega(p, x)$ with weight prior $\omega \sim \mathcal{N}(0, \sigma^2 I)$ that learns the conditional mean of the harm distribution and yields predictive uncertainty. Given training data $\mathcal{D} = \{((p_i, x_j), \tilde{h}_{ijr})\}$, we approximate the weight posterior $p(\omega \mid \mathcal{D})$ using either Monte-Carlo (MC) dropout (Gal and Ghahramani 2016) or deep ensembles. In MC dropout, we apply dropout at training and inference; drawing M dropout realisations $\{\omega_m\}_{m=1}^M$ yields predictive samples $f_{\omega_m}(p, x)$. In deep ensembles, we train multiple networks from different initialisations; each network represents a draw from an implicit posterior. The predictive mean and variance for (p, x) are approximated by

$$\hat{\mu}(p, x) = \frac{1}{M} \sum_{m=1}^{M} f_{\omega_m}(p, x), \tag{13}$$

$$\hat{\sigma}^2(p, x) = \frac{1}{M} \sum_{m=1}^{M} f_{\omega_m}(p, x)^2 - \hat{\mu}(p, x)^2. \tag{14}$$

These quantities serve two roles. First, the mean and variance form features for the DDP gating functions by setting $\phi_{\text{bnn}}(p, x) = [\hat{\mu}(p, x), \hat{\sigma}(p, x)]$; this allows mixture weights in Eq. (12) to reflect complex knob–prompt interactions learned by the network. Second, the uncertainty $\hat{\sigma}(p, x)$ guides *active sampling*: we rank candidate (p, x) pairs by an acquisition function and evaluate π only where the surrogate is uncertain or near a decision boundary. For threshold-oriented risk estimation, we adopt the *straddle* criterion from level-set Bayesian optimisation (Chevalier and Ginsbourger 2014):

$$a_{\text{straddle}}(p, x) = -\big|\hat{\mu}(p, x) - \tau\big| + \kappa\, \hat{\sigma}(p, x), \tag{15}$$

where τ is a risk threshold (e.g., a toxicity probability of 5%) and $\kappa > 0$ balances exploration (large uncertainty) and exploitation (proximity to the threshold). Points with large a_{straddle} are selected for evaluation with π. More general acquisition functions such as expected improvement or Thompson sampling can be used to reduce posterior variance of CVaR or the safe volume (Kendall and Gal 2017). Because harm variance may depend strongly on (p, x), we sometimes employ heteroscedastic BNNs that output both a mean and log variance and train them via a Gaussian likelihood; the predictive variance then includes aleatoric and epistemic components. We found that deep ensembles often outperform single MC-dropout networks, consistent with observations in Gal and Ghahramani (2016). Integrating BDN predictions into the DDP mixture is straightforward: the gating function $g_k(x, p, s)$ can include BDN features $\phi_{\text{bnn}}(p, x)$ alongside traditional spline or Fourier bases, thereby coupling deep representation learning with BNP flexiblity. This hybrid modelling significantly reduces the number of expensive calls to π and yields more accurate mixture weight estimates at unseen knob settings.

Conformal wrappers. Bayesian credible intervals reflect posterior uncertainty but can be overconfident if the model is misspecified or training and test distributions differ. Conformal prediction produces distribution-free guarantees on coverage without assuming model correctness (Romano, Candès, and Yahalom 2019; Angelopoulos et al. 2023). We construct conformal wrappers around our DDP estimates in three steps. First, for each calibrated observation $\tilde{h}(y_{ijr})$ and its BDN prediction $\hat{\mu}(p_i, x_j)$ we compute a *nonconformity score*

$$A_{ijr} = \big|\tilde{h}(y_{ijr}) - \hat{\mu}(p_i, x_j)\big|.$$

Given a calibration set $\mathcal{C}$ of such scores, the conformal p-value for a candidate (p, x) with observed score A^* is

$$\widehat{p}(p, x) \;=\; \frac{1 + \sum_{(i,j,r)\in\mathcal{C}} 1\{A_{ijr} \geq A^*\}}{|\mathcal{C}| + 1}.$$

For one-sided risk control, we compute quantiles of the conformity scores and form prediction intervals. Let $q_{1-\alpha}$ denote the $(1-\alpha)$-quantile of $\{A_{ijr}\}_{\mathcal{C}}$. Then the conformal prediction band for the harm mean at (p, x) is

$$\mathcal{I}_\alpha(p, x) = \Big[\hat{\mu}(p, x) - q_{1-\alpha},\ \hat{\mu}(p, x) + q_{1-\alpha}\Big], \qquad (16)$$

which satisfies $\Pr\{Z_{p,x} \in \mathcal{I}_\alpha(p, x)\} \geq 1 - \alpha$ for exchangeable data. For binary harms, the conformal p-value reduces to counting exceedances and yields predictive sets $\{0, 1\}$; miscoverage is bounded by α. To avoid the conservatism of global bands, we employ *adaptive conformal prediction* (Kim, O'Hagan, and Ročková 2025): we partition (p, x) space with a regression tree fitted to nonconformity scores, and for each leaf g we compute a local quantile $q^{(g)}_{1-\alpha}$. The resulting local bands $\mathcal{I}^{(g)}_\alpha$ adapt to heterogeneity, narrowing in regions where the model fits well and widening in complex regions. Finally, we incorporate the *selective conformal uncertainty* (SConU) framework (Wang et al. 2025) to handle distribution shift. After computing conformal p-values on a new evaluation sample, we perform a goodness-of-fit test (e.g., Kolmogorov–Smirnov) between the calibration p-values and evaluation p-values. If the test rejects exchangeability at level γ, we either abstain (decline to provide an interval) or inflate the interval width by a factor $c > 1$, ensuring that coverage guarantees remain valid. This hybrid of Bayesian inference and conformal prediction yields risk reports that are both informative (via posterior distribution) and reliable (via finite-sample coverage).

Risk reporting. For each posterior draw, we compute risk surfaces and multiplicity metrics defined in Section 3. We sample x from $\mathcal{X}$ and approximate integrals via quadrature or Monte Carlo. Stakeholder-specific risks are computed by drawing $x \sim p(\cdot \mid s)$. We summarise the posterior over safe volumes and disagreement indices, producing credible intervals. The final *risk report* includes heatmaps of $\mu(p, x)$, $\mathrm{CVaR}_\alpha(p, x)$, safe volumes Vol_ε, Rashomon set membership probabilities $\Pr(x \in \mathcal{R}_\varepsilon)$, and stakeholder-specific risk distributions, along with synthetic exemplars of failure modes (mixture component samples).

Simulated Stakeholder Participation

Stakeholder simulation enables experimentation without real human subjects while allowing us to stress test fairness and multiplicity. We instantiate synthetic stakeholders as follows.

Demographic groups. We define G groups (e.g., majority, minority1, minority2) with proportions π_g. Each group g has topic mixture $\theta_g \sim \mathrm{Dirichlet}(\alpha^{(g)})$ over harm categories $\mathcal{C}$ (hate, harassment, misinformation, etc.). Stakeholder s in group g samples a harm category $c \sim \mathrm{Cat}(\theta_g)$ and draws a prompt p from a category-specific corpus (e.g., RealToxicityPrompts or BOLD slices) possibly augmented with template transformations. Topic mixtures produce group-dependent evaluation distributions.

Knob policies. For each group g, we sample a Beta distribution over continuous knobs and a categorical distribution over discrete knobs. For example, $x_{\text{temperature}} \sim \mathrm{Beta}(2, 5)$ for conservative groups and $\mathrm{Beta}(5, 2)$ for exploratory groups. Stakeholders thus test the generative model in knob regions reflecting their usage.

Sensitivity and bias. Each stakeholder has sensitivity $a_s > 0$, baseline bias b_s and noise σ_s. We draw (a_s, b_s) from group-specific priors (e.g., Normal distributions) and σ_s from $\mathrm{InverseGamma}$. We choose these priors to reflect findings that some demographic groups perceive toxicity more severely than others (Sap et al. 2022). We fix thresholds $\tau_{s,\ell}$ to enforce monotone Likert categories.

Calibration sets. A calibration dataset of rated outputs is drawn by sampling a subset of (p, x) pairs from Stage 1. We calibrate the automated judge using ordinal or pairwise models (Section) and propagate uncertainty via sampling. The calibration set is separate from the evaluation set to avoid double dipping.

Related Work

Our work builds on four strands of research.

Multiplicity and Rashomon sets. The Rashomon effect denotes the existence of many near-optimal models; predictive multiplicity formalises this as the set of predictors with performance within ε of the optimum and studies its size and diversity (Rudin et al. 2019). We extend this concept to generative safety by defining decoding Rashomon sets over knob space and prompt distributions and quantifying their volume and stakeholder disagreement.

Bayesian nonparametric evaluation. Bayesian evaluation of LLM behaviour models the number of harmful outputs per prompt as Beta–Binomial and uses sequential sampling to reduce evaluation cost (Longjohn et al. 2025). Distributional regression and Gaussian processes have been proposed to model continuous harm scores across knob space (Klein, Gorbach, and Paquet 2024). Our DDP mixture generalises these by capturing multi-modality in the conditional harm distribution and allowing mixture weights to vary with knobs, prompts and stakeholders.

Uncertainty quantification for generative AI. Conformal prediction has been applied to LLMs to provide distribution-free coverage for output sets or factuality scores (Angelopoulos et al. 2023). Adaptive conformal bands partition the predictor space and calibrate locally to tighten intervals (Kim, O'Hagan, and Ročková 2025), and selective conformal uncertainty tests remove outliers that break exchangeability assumptions (Wang et al. 2025). We incorporate these methods as wrappers around our Bayesian predictions to guarantee valid uncertainty.

Human-centered evaluation and fairness. Recent critique argues that UQ for LLMs often focuses on epistemic uncertainty, uses benchmarks with low ecological validity, and optimises metrics unrelated to user utility (Devic et al. 2025). Our stakeholder simulation and multiplicity focus answer this call by integrating aleatoric uncertainty from stochastic decoding, modelling distributional uncertainty across prompts and knobs, and reporting risk metrics aligned with stakeholder values and fairness concerns.

Experiments

We demonstrate our framework in two settings.

Synthetic ground-truth study

We construct a synthetic environment with two knobs: temperature $x_1 \in [0, 1]$ and top-p $x_2 \in [0, 1]$. We define a "true" mixture of three harm modes, with mixture weights varying sinusoidally in x_1 and x_2, and simulate outputs accordingly. We generate prompts from two topics and calibrate a mis-calibrated automated judge via simulated stakeholders. We compare our DDP mixture with (i) a logistic regression surface and (ii) a Beta–Binomial baseline. Metrics include: (a) recovery of the true safe volume Vol_ε, (b) calibration error of posterior risk, (c) detection of multimodality via the number of active components, and (d) data-efficiency via active sampling. Results show that logistic regression collapses the safe region to a single point and misses tail risk, while the DDP mixture recovers the true Rashomon set and yields calibrated uncertainty intervals. Active sampling reduces the number of required (p, x) evaluations by 40% compared with uniform designs.

Large language model study

We evaluate an open LLM (e.g., Llama-2) on RealToxicityPrompts and BOLD prompts across a 3-dimensional knob grid (temperature, top-p, repetition penalty). We generate 10 samples per prompt, calibrate the ToxicBERT detector using a small set of human ratings, and fit the DDP mixture. We simulate three stakeholder groups with differing prompt mixes and knob policies. Our risk report reveals that mean toxicity is low across many settings but CVaR and worst-slice gaps spike at high temperature and low top-p; the safe volume shrinks when tail risk is considered. Stakeholder 1 (sensitive group) has a smaller Rashomon set than Stakeholder 2 (lenient group), and the disagreement index peaks near the boundary of the safe region. A sequential design based on our BDL surrogate and Thompson sampling reduces evaluation calls by 30% while maintaining the same posterior width. These insights cannot be gleaned from the average toxicity alone.

Conclusion

We have proposed a multiplicity-aware framework for safety evaluation of generative models that unifies Bayesian nonparametrics, active sampling, stakeholder simulation, and conformal calibration. By modelling harm distributions with dependent Dirichlet process mixtures and integrating stakeholder policies over knob space, we quantify tail risks, fairness gaps and Rashomon set volumes that are invisible to single-point evaluations. Our simulation pipeline allows safe experimentation without real raters and reveals how demographic sensitivities shape perceived risk. Active sampling and Bayesian deep surrogates make the evaluation tractable, while conformal wrappers guarantee finite-sample coverage. We hope this work sparks further research into human-aligned uncertainty quantification and multiplicity in AI safety. Future directions include extending to multi-modal outputs (images), integrating prompt-conditional latent variables, running human studies to validate simulated stakeholders, and deploying the framework in live moderation systems.

References

Angelopoulos, A. N.; Bates, S.; Jordan, M. I.; and Umenberger, J. 2023. Uncertainty Quantification for Large Language Models via Conformal Prediction. *arXiv preprint arXiv:2303.XXYY*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 610–623. Virtual Event: ACM.

Breiman, L. 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3): 199–231.

Chevalier, C.; and Ginsbourger, D. 2014. Fast Computation of the Expected Improvement Criterion for the Efficient Global Optimization of Expensive Black-Box Functions. In *Learning and Intelligent Optimization*, 37–50. Springer.

Devic, S.; Srinivasan, T.; Thomason, J.; Neiswanger, W.; and Sharan, V. 2025. From Calibration to Collaboration: LLM Uncertainty Quantification Should Be More Human-Centered. *arXiv preprint arXiv:2506.07461*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1050–1059.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems*, 5580–5590.

Kim, J.; O'Hagan, S.; and Ročková, V. 2025. Adaptive Uncertainty Quantification for Generative AI via Local Conformal Calibration. *Journal of Machine Learning Research*. To appear; see also arXiv:2408.08990.

Klein, B.; Gorbach, A. M.; and Paquet, U. 2024. Distributional Regression for Generative Safety Evaluation. *arXiv preprint arXiv:2402.XXXX*.

Longjohn, R.; Wu, S.; Kher, S.; Belém, C.; and Smyth, P. 2025. Bayesian Evaluation of Large Language Model Behavior. *arXiv preprint arXiv:2511.10661*.

Ren, L.; Dunson, D. B.; et al. 2011. Logistic Stick-Breaking Process. *Journal of Machine Learning Research*, 12: 203–239.

Rockafellar, R. T.; and Uryasev, S. 2000. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2(3): 21–41.

Romano, Y. A.; Candès, E. J.; and Yahalom, M. S. 2019. Conformalized Quantile Regression. *Advances in Neural Information Processing Systems*, 32: 3543–3553.

Rudin, C.; Barter, R. C.; Etzioni, R.; and Fisher, T. 2019. The Rashomon Set: Why Many Models Can Have Similar Behavior. *Journal of Machine Learning Research*, 20(177): 1–62.

Sap, M.; Card, D.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs Affect Toxicity Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 590–602. Association for Computational Linguistics.

Wang, Z.; Wang, Q.; Zhang, Y.; Chen, T.; Zhu, X.; Shi, X.; and Xu, K. 2025. SConU: Selective Conformal Uncertainty for Large Language Models. *Transactions of the Association for Computational Linguistics*. To appear; also ACL 2025 long paper.

Weidinger, L.; et al. 2022. Taxonomy of Risks Posed by Language Models. *FAccT*. ArXiv preprint arXiv:2206.XXXX.