"Do Your Guardrails Even Guard?" Method for Evaluating Effectiveness of Moderation Guardrails in Aligning LLM Outputs with Expert User Expectations

Anindya Das Antar, Xun Huan, Nikola Banovic

University of Michigan, Ann Arbor
adantar@umich.edu, xhuan@umich.edu, nbanovic@umich.edu

Abstract

Ensuring that large language models (LLMs) align with human values and goals is crucial for their adoption in high-stakes decision-making. To guard against incorrect, misleading, or otherwise unexpected or undesirable LLM outputs, guardrail engineers implement guardrails based on expert knowledge from subject-matter authorities to steer and align pre-trained LLMs. Existing evaluation methods assess LLM performance, with and without guardrails, but provide limited insight into the contribution of each individual guardrail and its interactions on alignment. Here, we present a method to evaluate and select guardrails that best align LLM outputs with empirical evidence representing expert knowledge. Through evaluation with real-world illustrative examples of resume quality and recidivism prediction, we show that our method effectively identifies useful moderation guardrails in a way that could help guardrail engineers interpret contributions of different guardrails to "user-LLM" alignment.

Introduction

Ensuring that LLMs produce desirable outputs without harmful side-effects (Terry et al. 2024) (i.e., that they achieve "user-LLM alignment") is crucial for their broader adoption. This includes aligning LLM outputs with user intentions (Boggust et al. 2022), organizational goals (Deshpande and Sharp 2022), existing domain knowledge (Zhou et al. 2024), empirical evidence (Wang et al. 2024a), societal values (Arzberger et al. 2025; Estrada 2018), and ethical considerations (Shneiderman 2020; Norhashim and Hahn 2024). However, despite training on vast data, LLMs can still produce incorrect, misleading, or otherwise unexpected or undesirable outputs (Bender et al. 2021; Dutta et al. 2024).

Existing research (Rebedea et al. 2023; Inan et al. 2023) showed that guardrails can steer pre-trained LLM outputs and shape decision-making. Moderation guardrails (Mahomed et al. 2024), implemented through prompt engineering, translate expert input, corporate guidelines, and empirical findings into model behavior by injecting constraints, filtering criteria, or value-aligned instructions into prompts. Compared to costly, resource-intensive LLM retraining, moderation guardrails provide a lightweight, flexible way to align off-the-shelf LLMs with user expectations.

Yet, not all guardrails are effective; some even introduce biases that negatively steer and impact LLM decision-making (Chu et al. 2024). Thus, it is essential to identify and retain only those guardrails that enhance alignment. However, existing LLM evaluation methods (Chang et al. 2024) do not specifically focus on evaluating and selecting individual guardrails from a set to ensure user-LLM alignment.

Here, we present a probabilistic method using Bayesian inference to evaluate the effectiveness of guardrails in aligning a pre-trained LLM with expert user expectations for decision-making (Fig. 1). We answer the following research question: How can guardrail engineers systematically evaluate and select guardrails to ensure LLM outputs align with expert user expectations? We do so by producing measurable and interpretable estimates of the contribution of individual guardrails and their interactions on model alignment.

Our method uses a curated empirical alignment dataset $\mathcal{D}$ reflecting expert expectations for LLM outputs to evaluate a set of moderation guardrails $\mathcal{G}$ meant to align the LLM with those expectations. Our method then estimates a vector of Bernoulli parameters θ, representing the activation probability of each corresponding guardrail in $\mathcal{G}$, and models the posterior probability distribution of θ to capture uncertainty in its estimation. Unlike deterministic methods that simply toggle guardrails on or off, our approach treats θ as a continuous probability, allowing for a more nuanced evaluation.

We illustrated and validated our method in two realistic use cases: 1) resume quality classification (Bertrand and Mullainathan 2004) and 2) recidivism prediction (Angwin et al. 2016). Each use case uses knowledge from subject matter authority experts along with existing empirical data to construct an alignment dataset $\mathcal{D}$ and a set of moderation guardrails $\mathcal{G}$, including specially designed distractor guardrails that should not be selected. Our results showed that our method selects guardrails that align LLM predictions with expert expectations better than the model alone.

Our method systematically evaluates and selects moderation guardrails, enabling guardrail engineers to align LLM decisions with expert user expectations. By quantifying uncertainty in guardrail selection, our method helps them interpret the confidence of probabilistically activating each guardrail. Such rigorous and nuanced evaluation has the potential to ensure responsible deployments of LLMs in a way that increases their adoption in high-stakes decision-making.

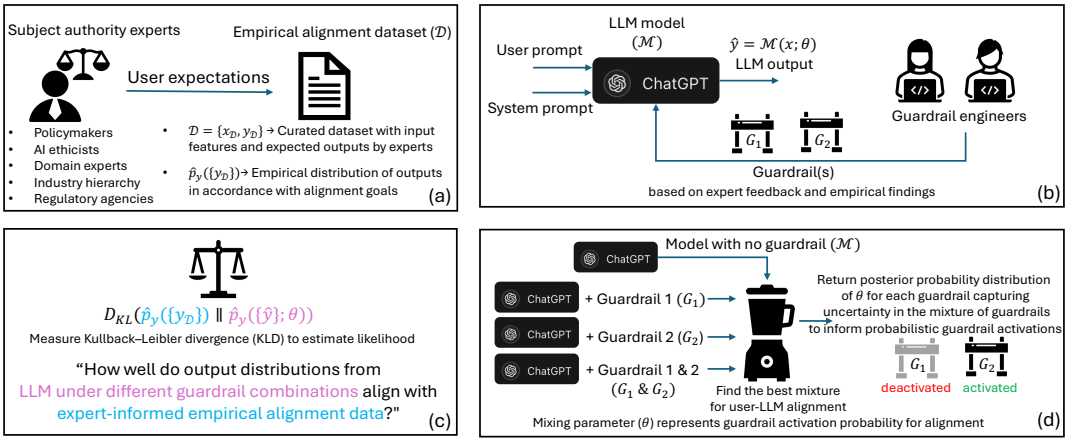

Figure 1: Our method for evaluating and selecting guardrails: (a) empirical alignment dataset reflecting expert user expectations, (b) implementation of expert-informed guardrails by guardrail engineers, (c) measuring alignment between outputs from the LLM and expert-informed empirical distribution, and (d) identifying effective guardrail combinations to enhance alignment.

Background and Related Work

Here, we describe the goals of aligning LLM outputs with user expectations, explain how different guardrails steer model outputs to achieve this alignment, and review existing guardrail evaluation methods, highlighting their limitations.

Aligning Model Outputs to User Expectations

Pre-trained LLM outputs are often misaligned with user expectations due to their inherent randomness and presence of biases (Bender et al. 2021; Chang et al. 2024). Aligning LLM outputs with user expectations (Shankar et al. 2024; Hrytsyna and Alves 2024), rules (Hadfield-Menell, Andrus, and Hadfield 2019), and values (Norhashim and Hahn 2025; Gabriel 2020) is essential for accurate decision-making (Lai et al. 2023; Chiang et al. 2024) and mitigating bias (Fang et al. 2023). Effective user-LLM alignment (Wang et al. 2024b; Gombolay 2024; Robinson 2023) goes beyond technical adjustments—it requires understanding the stakeholders who define these expectations and shape model behavior (Deshpande and Sharp 2022). Domain experts, ethicists, and policymakers establish alignment objectives based on domain knowledge (Zhi-Xuan et al. 2024; Zhang et al. 2024), empirical data (Wang et al. 2024a), societal values (Norhashim and Hahn 2024; Varshney 2025), law (Hadfield-Menell and Hadfield 2019), and ethical guidelines (Chivukula et al. 2024; Awad et al. 2018). Guardrail engineers operationalize these objectives by identifying risks, implementing mitigations, and shaping alignment strategies in collaboration with domain experts (Chiang et al. 2023).

Influence of Guardrails in Model Alignment

Existing human-AI alignment methods integrate explicit principles during LLM development (Gabriel 2020) or fine-tune models with high-quality data (Zhao et al. 2024). However, unlike narrow AI models, retraining general-purpose LLMs is costly (Wang et al. 2024b). Guardrails provide an alternative by steering pre-trained LLM outputs (Dong et al. 2024). Input guardrails prevent inappropriate content from reaching the model, while output guardrails validate and refine responses before delivery. Output moderation guardrails (Yuan et al. 2024) enforce corporate policies and ethical guidelines to align LLM decisions with user expectations. Also, security guardrails help mitigate jailbreaking (Wang et al. 2024c; Cornacchia et al. 2025) and adversarial attacks (Dev et al. 2024; Iqbal, Kohno, and Roesner 2025), both of which are beyond the scope of this paper.

Existing Approaches to Guardrail Evaluation

Evaluating and selecting effective guardrails remains challenging due to the difficulty of isolating their individual effects and understanding their interplay. Existing LLM evaluation methods that assess output quality (Liu et al. 2023) or fairness (Li et al. 2020) could, in principle, be repurposed to measure changes in LLM behavior after applying guardrails. However, this could obscure the contributions of individual guardrails and their interactions. Methods that steer LLMs by manipulating internal representations (e.g., changing activations) (Chu et al. 2024) could highlight discrepancies between generated and desirable outputs, but attempt to modify the model rather than prompt-based guardrails. Existing auditing methods (Mahomed et al. 2024; Rastogi et al. 2023) evaluate which LLM outputs are affected by existing guardrails, but not which guardrails meaningfully contribute to improved alignment. This limits their usefulness for guardrail selection, even when combined with methods that detect specific guardrail presence (Yang et al. 2025).

Method to Evaluate and Select Guardrails

Our probabilistic method leverages Bayesian inference to evaluate and select guardrails that best align pre-trained LLM outputs with expert expectations. By modeling each guardrail's posterior probability distribution, our method helps interpret the uncertainty of activating each guardrail. It also helps identify and exclude guardrails that could introduce unintended downstream consequences in predictions.

User-Defined Inputs for Guardrail Evaluation

The guardrail engineer using our method (henceforth *the user*) must define alignment criteria for a pre-trained LLM $\mathcal{M}$ to align its outputs with a curated dataset $\mathcal{D}$ that contains empirical evidence representing expert knowledge. The user may adopt a set of guardrails $\mathcal{G}$ from existing relevant literature or design new ones informed by expert feedback.

Empirical Alignment Data $\mathcal{D}$: To align LLM behavior with expert expectations, the user must define an alignment dataset $\mathcal{D} = \{x_{\mathcal{D}}, y_{\mathcal{D}}\}$, where $x_{\mathcal{D}}$ encodes prompt features (e.g., building blocks capturing role, task, requirements, instructions in a prompt) and $y_{\mathcal{D}}$ represents expert-informed ground-truth outcomes (Fig. 1a). Subject-matter authorities, such as policymakers, AI ethicists, or domain experts, define these outcomes using domain knowledge, empirical evidence, and ethical considerations. To ensure $\mathcal{D}$ reflects expert knowledge, ground-truth labels should come directly from experts or represent real-world outcomes validated by experts. The alignment dataset $\mathcal{D}$ should also promote equal opportunity and fairness (e.g., it should be demographically balanced). Users may consult experts to create and curate such datasets or select existing ones. Our method assumes $\mathcal{D}$ accurately reflects the intended alignment outcomes, though verifying its correctness is beyond this paper's scope.

LLM $\mathcal{M}$: The user must select a pre-trained off-the-shelf LLM (e.g., GPT, PaLM, Gemini, LLaMA, Claude, FLAN-T5) (Chang et al. 2024). Although our model-agnostic method works with any LLM, the guardrails it evaluated and selected for one LLM may not generalize across models due to variations in architecture, training data, and sensitivity to different guardrails. We define an outcome $\hat{y} = \mathcal{M}(x; \theta)$ that encodes the prediction output from an LLM $\mathcal{M}$ for prompt constructed from prompt features x, given guardrail activation probability vector θ (Fig. 1b).

System Prompts and User Prompts: To probe the LLM, the user must define system prompts and craft user prompts (Fig. 1b). System prompts specify guardrail constraints, including the model's role, task, requirements, etc. In addition to guardrails for evaluation, the system prompt may contain other instructions that should be fixed across all queries to the LLM to isolate the effect of the guardrails. For example, such information could assign the LLM a consistent persona (e.g., "You are a resume evaluator") to standardize its output (Zheng et al. 2024). User prompts contain tailored inputs from the alignment data $\mathcal{D}$ to elicit LLM responses.

Guardrails $\mathcal{G}$ for Evaluation: The user must adopt a set of guardrails, $\mathcal{G} = \{G_1, G_2, G_3, \ldots, G_N\}$ for evaluation. They can choose these guardrails by consulting existing subject matter experts, reviewing literature, or implementing new ones by identifying LLM limitations (Fig. 1b). These guardrails may include alignment constraints, counter-bias measures, etc. Our preliminary experiments suggested that guardrail order (e.g., G_i preceding G_j, or G_j preceding G_i) can influence LLM outputs. However, we found no conclusive evidence that LLMs systematically prioritize guardrails by order. Nevertheless, to mitigate potential order effects, we randomize guardrail sequences in our method.

Task-specific Surrogate Model to Mimic LLM

Querying an LLM for every record across all guardrail combinations is computationally expensive due to token pricing, request limits, and latency, with costs scaling as $2^N \times R$ for R records and N guardrails. To reduce this, users can *optionally* train a surrogate Deep Neural Network (DNN) (Shree et al. 2024; Pangakis and Wolken 2024) that predicts task-specific outcomes from encoded prompt features, enabling faster and cheaper per-query evaluation. This choice is driven by online LLM query cost and latency rather than training cost, as DNNs also incur some overhead. If LLMs run locally with low latency and no token pricing, or if online LLM cost and query time are not concerns, our method can be applied directly without a surrogate DNN.

Need For Surrogate Model: We emphasize that training a surrogate model is optional for accelerating queries, and *not* a methodological requirement for selecting effective guardrails. Our method is model-agnostic and can directly work with offline LLMs with low latency and no token pricing, or when an online LLM query cost is manageable. Also, we do not claim that the surrogate approximates the full LLM capabilities, as it only mimics LLM's task-specific behavior (e.g., prediction) under controlled prompting. Although we train the surrogate for prediction tasks in this paper, the same approach can extend to other tasks (e.g., text generation), depending on the nature of the LLM outputs and how well they can be structured for supervised learning.

Input Data Preprocessing: The user must preprocess the alignment dataset $\mathcal{D}$ to create separate training and test sets with structured feature representations suitable for the surrogate model. While LLMs operate on natural language prompt features x derived from the raw records $x_{\mathcal{D}} \subset \mathcal{D}$, DNNs require standardized, encoded inputs x^{encoded}. Thus, the user must transform the raw records into encoded features (e.g., by one-hot encoding categorical variables and normalizing continuous variables to a fixed range, such as $[0, 1]$). Missing values can be addressed either by discarding affected records or by imputation, using strategies such as mean, median, or K-nearest neighbors. Finally, each encoded input vector must be paired with a binary guardrail activation indicator vector $\mathbf{g} \in \{0, 1\}^N$ indicating which guardrails were activated during the original LLM query using the corresponding system and user prompt.

Ground Truth: Since the surrogate DNN is intended to mimic the LLM's task-specific behavior (e.g., prediction) across different guardrail combinations, it must be trained on structured LLM-generated outcomes rather than expert-labeled ground truth from $\mathcal{D}$. To obtain these, the user should query the LLM using the original prompt features and corresponding guardrails aligned with the encoded DNN inputs. Based on the task, each LLM response should then be mapped to a structured outcome label (e.g., binary or multi-class) using task-specific keyword matching, regular expressions, or other postprocessing heuristics. These label mappings must be applied consistently across records and be robust to variation in LLM outputs. The resulting structured labels serve as ground truth for training the DNN surrogate.

Surrogate Model Training and Testing: The DNN's training and testing sets should capture diverse guardrail activation patterns. Having R raw records (e.g., selected among inputs $x_{\mathcal{D}} \subset \mathcal{D}$) and N guardrails results in $2^N \times R$ prompt–guardrail combinations. This is typically too large and prohibitively expensive to query exhaustively from an online LLM just to generate DNN training labels.

To address this, users can instead sample a representative subset $\tilde{\mathcal{D}} = \{(x_i, \mathbf{g}_i)\} \subseteq \mathcal{D} \times \{0,1\}^N$ from all possible combinations of prompts and guardrail activation indicator vectors using random or stratified sampling. Stratified sampling helps ensure sufficient representation across guardrail configurations. The collected samples can then be split into training, validation, and testing (e.g., 60%/10%/30% split).

Users can tune the surrogate DNN. For example, to validate our method, we tuned our surrogate DNNs using grid-search cross-validation with early stopping, varying layers, neurons, learning rates, activations, batch sizes, dropout, and optimizers (see "Validation and Results"). We chose grid search for its consistency and search coverage. Users may also use automated tuning tools, which may yield slightly different DNN parameters without affecting evaluation.

Modeling Guardrail Activation

We model guardrail activation probabilities as a vector of Bernoulli parameters θ, with its distribution capturing uncertainty in guardrail activation. Sampling from it can help identify guardrail combinations that improve alignment.

Empirical Distribution of Outputs $\hat{p}_y(\{y_{\mathcal{D}}\})$: We construct an empirical distribution of outcomes $\hat{p}_y(\{y_{\mathcal{D}}\})$ from all outputs $\{y_{\mathcal{D}}\}$ of the alignment dataset $\mathcal{D}$ (Fig. 1a). This serves as the target distribution for aligning the LLM's output to identify the optimal combination of guardrails and will be used later to define the Bayesian likelihood.

Guardrail Activation Probability θ: We represent the stochastic activation of each guardrail using a probability vector $\theta = [\theta_1, \theta_2, \ldots, \theta_N]$, where each θ_i represents the activation probability of guardrail G_i when prompting the LLM for a specific task (e.g., binary classification):

$$\theta_i = P(\mathbf{g}_i = 1) = 1 - P(\mathbf{g}_i = 0), \tag{1}$$

where $\mathbf{g}_i$ is a binary guardrail activation indicator vector, with $\mathbf{g}_i = 1$ when guardrail G_i is active or $\mathbf{g}_i = 0$ otherwise. Hence, each guardrail is activated through a biased coin flip with activation probability θ_i simulated by Bernoulli trials.

Prior Distribution $P(\theta)$: The prior distribution $P(\theta)$ characterizes the initial uncertainty in θ before seeing any data. We select a truncated normal prior that favors θ values closer to 0.5; i.e., centering around a fair coin where each guardrail is equally likely to be activated or deactivated:

$$P(\theta) \sim \mathcal{TN}(\theta; \mu = 0.5, \sigma = 0.5, [0,1]), \tag{2}$$

where each θ_i is truncated to the interval $[0,1]$. Our prior is not tailored to the alignment dataset; it reflects a belief in the initial neutrality of the guardrails. This flexible prior allows the posterior to meaningfully update based on evidence from the alignment dataset without being overly restrictive.

For robustness, we tested three priors: 1) a truncated Laplace prior centered at 0.5 to emphasize central mass, 2) a uniform prior representing maximum initial uncertainty over θ, and 3) a wide Gaussian prior ($\mu = 0.5$, $\sigma = 2$), reflecting high but not maximal uncertainty about θ with a slight initial preference to neutrality. Across all priors, posterior inference produces consistent results: posterior means $\gg 0.5$ trigger stronger guardrail activation, $\ll 0.5$ to stronger deactivation, and ≈ 0.5 indicates low/no impact or indifference, suggesting our method is robust to reasonable prior choices.

Bayesian Likelihood $P(\{y_{\mathcal{D}}\}|\theta)$: We define the likelihood based on the empirical output distribution constructed from the ensemble of samples in the alignment dataset $\mathcal{D}$:

$$P(\{y_{\mathcal{D}}\}|\theta) \propto \frac{1}{D_{\mathrm{KL}}(\hat{p}_y(\{y_{\mathcal{D}}\}) \parallel \hat{p}_y(\{\hat{y}\};\theta))}, \tag{3}$$

where $\hat{p}_y(\{\hat{y}\};\theta)$ is similar to the desired empirical distribution of outputs $\hat{p}_y(\{y_{\mathcal{D}}\})$ except now constructed from LLM predictions $\{\hat{y}\} = \mathcal{M}(\{x\};\theta)$, where prompts $\{x\}$ are derived from inputs $\{x_{\mathcal{D}}\}$ of the alignment dataset $\mathcal{D}$, and D_{KL} is the Kullback–Leibler divergence (KLD) (Fig. 1c) (Kullback and Leibler 1951). Since a lower KLD implies more similar distributions, we use the inverse of KLD as a proxy for Bayesian likelihood estimation (MacKay 2002; Kato, Imaizumi, and Minami 2023). Thus, a $\hat{p}_y(\{\hat{y}\})$ more similar to that from the alignment data $\hat{p}_y(\{y_D\})$ yields a lower KLD, and in turn higher likelihood. This choice reflects our objective to evaluate which guardrails bring the model's behavior closer to the expert-informed output distribution.

Posterior Distribution $P(\theta|\{y_{\mathcal{D}}\})$: Finally, we obtain the posterior distribution capturing the updated uncertainty in θ after conditioning on the alignment dataset via Bayes' rule:

$$P(\theta|\{y_{\mathcal{D}}\}) = \frac{P(\{y_{\mathcal{D}}\}|\theta)P(\theta)}{P(\{y_{\mathcal{D}}\})}, \tag{4}$$

where all terms are also dependent on $\{x_D\}$, but omitted for brevity. We use the Metropolis–Hastings Markov chain Monte Carlo (MCMC) algorithm, implemented in the PyMC3 library, to generate 10,000 samples from the posterior distribution with 2,500 burn-in steps (Algorithm 1). Each θ sample acts as a realization of the Bernoulli parameter, guiding a biased coin flip to determine whether the corresponding guardrail is activated for each LLM query (Fig. 1d). The user can also evaluate the distributional properties of the posterior (e.g., mean, variance, quantiles) to interpret each guardrail's activation confidence and its uncertainty. For example, a posterior mean of θ deviating significantly from 0.5 indicates a stronger probability of choosing a biased coin—favoring either activation (mean $\gg 0.5$) or deactivation (mean $\ll 0.5$) of the corresponding guardrail.

Guardrail Selection

We summarize the guardrail selection process in Algorithm 2, using learned posterior distributions of θ. We propose three alternative strategies for selecting which guardrails to activate when predicting LLM outputs $\{\hat{y}\}$ from prompt features $\{x\}$ derived from raw records $\{x_{\mathcal{D}}\}$ of

Algorithm 1: Metropolis–Hastings MCMC for sampling the Bayesian posterior of guardrail activation parameter θ

Input: Empirical alignment dataset $\mathcal{D} = \{x_\mathcal{D}, y_\mathcal{D}\}$, prior distribution $P(\theta)$ for guardrail activation
Parameter: Number of guardrails N, number of samples S, number of chains C, burn-in steps B
Output: MCMC samples from the posterior distribution $P(\theta|\{y_\mathcal{D}\})$ after burn-in

Compute desired output distribution $\hat{p}_y(\{y_\mathcal{D}\})$ from the empirical alignment data $\mathcal{D}$ for all records $\{x_\mathcal{D}\}$
for each chain $c = 1$ to C **do**
  Initialize $\theta^{(0)} = [\theta_1^{(0)}, \ldots, \theta_N^{(0)}]$ by jointly sampling from prior $P(\theta)$ for all N guardrails
  for each iteration $s = 1$ to $S + B$ **do**
    Propose new $\theta' = [\theta'_1, \ldots, \theta'_N]$ from proposal distribution (e.g., Gaussian centered at current chain location $\theta^{(s-1)}$)
    for each guardrail $i = 1$ to N **do**
      Perform Bernoulli coin toss with θ'_i as success probability to activate or deactivate guardrail G_i
    end for
    With activated guardrails generate LLM outputs $\{\hat{y}\} = \mathcal{M}(\{x\}; \theta')$ using prompts $\{x\}$ from $\{x_\mathcal{D}\}$
    Generate LLM output distribution with current guardrail activation: $\hat{p}_y(\{\hat{y}\}; \theta')$
    Compute KLD between $\hat{p}_y(\{y_\mathcal{D}\})$ and $\hat{p}_y(\{\hat{y}\}; \theta')$ and approximately estimate likelihood $P(\{y_\mathcal{D}\}|\theta')$
    Compute acceptance probability:

$$\alpha = \min\left(1, \frac{P(\{y_\mathcal{D}\}|\theta')P(\theta')}{P(\{y_\mathcal{D}\}|\theta^{(s-1)})P(\theta^{(s-1)})}\right)$$

    Accept θ' with probability α; otherwise, reject by setting $\theta^{(s)} = \theta^{(s-1)}$
  end for
  Discard first B samples from each chain as burn-in
end for
return Retain $\{\theta^{(s)}\}$ after burn-in from all chains collectively as samples from the posterior $P(\theta|\{y_\mathcal{D}\})$

Algorithm 2: Guardrail selection using different activation strategies to enhance user-LLM alignment

Input: Alignment dataset $\mathcal{D} = \{x_\mathcal{D}, y_\mathcal{D}\}$ for evaluation; posterior distribution $P(\theta|\{y_\mathcal{D}\})$ over all guardrails
Parameter: Number of guardrails N, activation `strategy` $\in$ {`bernoulli_toss`, `mean_cutoff`, `sample_cutoff`}
Output: Selected guardrail vectors $\mathcal{G}$ and corresponding LLM outputs $\hat{\mathcal{Y}}$ for all records in $\mathcal{D}$

if `strategy` $=$ `mean_cutoff` **then**
  For each guardrail G_i, compute posterior mean $\bar{\theta}_i = \mathbb{E}[\theta_i]$ from the posterior distribution $P(\theta_i|\{y_\mathcal{D}\})$
end if
Initialize $\mathcal{G} = [\,]$ {Stores selected guardrail vectors}
Initialize $\hat{\mathcal{Y}} = [\,]$ {Stores corresponding LLM outputs}
for each record $(x_\mathcal{D}, y_\mathcal{D}) \in \mathcal{D}$ **do**
  Initialize binary guardrail vector $\mathbf{g} = [\,]$
  if `strategy` $\in$ {`bernoulli_toss`, `sample_cutoff`} **then**
    Sample joint vector $\hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_N]$ from the posterior $P(\theta|\{y_\mathcal{D}\})$ {Sample jointly to preserve posterior correlations among N guardrail activations}
  end if
  for each guardrail G_i, where $i = 1, \ldots, N$ **do**
    if `strategy` $=$ `bernoulli_toss` **then**
      Perform Bernoulli coin toss with success probability $\hat{\theta}_i$ {similar to tossing a fair or biased coin}
      Set $\mathbf{g}_i = 1$ if toss is heads, else $\mathbf{g}_i = 0$
    else if `strategy` $=$ `sample_cutoff` **then**
      Set $\mathbf{g}_i = 1$ if $\hat{\theta}_i > 0.5$, else $\mathbf{g}_i = 0$
    else if `strategy` $=$ `mean_cutoff` **then**
      Set $\mathbf{g}_i = 1$ if $\bar{\theta}_i > 0.5$, else $\mathbf{g}_i = 0$
    end if
    Append $\mathbf{g}_i$ to $\mathbf{g}$
  end for
  Build prompt x using $x_\mathcal{D}$ and $\mathbf{g}$ to get LLM output $\hat{y}$
  Append $\mathbf{g}$ to $\mathcal{G}$ and $\hat{y}$ to $\hat{\mathcal{Y}}$
end for
return $\mathcal{G}$ and $\hat{\mathcal{Y}}$

the empirical alignment dataset $\mathcal{D}$. The user can then compare resulting LLM outputs $\{\hat{y}\}$ against expert-informed ground truth labels $\{y_\mathcal{D}\}$, using a task-specific evaluation metric of their choice to assess user-LLM alignment. When $\mathcal{D}$ is already curated to ensure fairness (e.g., through demographic balancing) and has expert-informed labels, predictive performance metrics (e.g., F1 score (Rijsbergen 1979)) can be used to evaluate how closely $\{\hat{y}\}$ align with $\{y_\mathcal{D}\}$. If not, fairness metrics (e.g., equalized odds, demographic parity) (Hardt, Price, and Srebro 2016) may be better suited to assess alignment through an equity lens.

In the *Bernoulli toss* and *sample cut-off* strategies, the user samples a joint vector $\hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_N]$ from the posterior $P(\theta|\{y_\mathcal{D}\})$, where each $\hat{\theta}_i$ represents the activation probability of guardrail G_i, preserving posterior correlations. In the Bernoulli toss, $\hat{\theta}_i$ is used as the success probability in a coin toss to decide whether to activate guardrail G_i. In the sample cut-off, $\hat{\theta}_i$ is compared directly to a threshold (typically 0.5): the guardrail is activated if $\hat{\theta}_i > 0.5$.

In the *mean cut-off* strategy, the user does not need to sample. Instead, they precompute the posterior mean $\bar{\theta}_i = \mathbb{E}[\theta_i]$ for each guardrail G_i and activate the guardrail if $\bar{\theta}_i > 0.5$. This approach uses the expected value of the posterior, either computed analytically if parameters are known or by fitting a parametric distribution to MCMC samples.

Limitations

Although our guardrail selection method provides a systematic framework for aligning LLM outputs with user expectations, it has limitations. If the alignment dataset $\mathcal{D}$ is biased, incomplete, or poorly curated, our method will still function, but resulting alignments will reflect those flaws. As with any alignment method relying on gold-standard ground truths, creating high-quality expert-informed datasets requires sig-

nificant effort. Since preparing such data is beyond this paper's scope, we used publicly available datasets already curated with expert input (e.g., professional resume assessments, real recidivism outcomes). Where such data is unavailable, we recommend involving domain experts and following established data curation best practices.

While the posterior distribution of θ provides insights into each guardrail's effectiveness and uncertainty, users without a statistics background may struggle with probabilistic data and visualizations (Prabhudesai et al. 2023). Here, we assume that our target users (i.e., guardrail engineers) can interpret uncertainty from probability distributions. Also, we used our method to evaluate guardrails globally to determine which are accepted or rejected for alignment, but we did not apply it to analyze local (individual input–output) or cohort-level (subgroup) influences on LLM outputs.

Three authors, two Computer Science and Engineering researchers and one Mechanical Engineering researcher, conducted quantitative evaluations. Two external researchers informed dataset selection and alignment criteria, but we did not conduct a qualitative evaluation with domain experts (e.g., resume reviewers or criminal justice experts), whose input could further strengthen user–LLM alignment.

Using the surrogate model consistently across guardrail configurations reduces LLM approximation bias and accelerates evaluation due to low DNN query costs. However, our model-agnostic method can also run on offline LLMs without token costs or online LLMs when latency and pricing allow. To manage combinatorial explosions, users can batch guardrails by related alignment goals or strong interactions, reducing evaluation burden while maintaining flexibility.

We selected moderation guardrails to evaluate our method in two high-stakes tasks (resume quality and recidivism prediction), where user-LLM alignment is critical. While our method generalizes to other guardrail types (e.g., jailbreak prevention, hallucination reduction) and goals (e.g., safety, security), doing so requires domain-specific data and alignment criteria. In domains with contested goals, such as political discourse or policy moderation, the main challenge is defining alignment targets, not applying our method.

Validation and Results

We conducted a series of quantitative experiments to validate our method's ability to select effective and reject ineffective guardrails for user-LLM alignment. Acting as guardrail engineers, we evaluated our method using two publicly available datasets in two real-world applications: 1) aligning LLM decisions with expert knowledge to promote equal opportunity in hiring (Bertrand and Mullainathan 2004), and 2) fairness in criminal justice (Angwin et al. 2016). We also consulted two researchers in computational law and algorithmic accountability to select datasets, define alignment criteria, and review initial guardrails, though they did not label ground truths or serve as co-authors.

Our method provides a Bayesian, quantitative approach to evaluating and selecting effective guardrails. Instead of binary on/off decisions, our posteriors yield probability distributions capturing uncertainty and evidence strength, guiding task-specific guardrail selection or rejection.

Validation LLM

To validate our method, we used U-M GPT, our institutional LLM with API access to GPT-4o. Although we used the same LLM in both validation tasks, we varied its role (e.g., resume quality or recidivism predictor) and guardrails via system prompts for task-specific features (e.g., candidate work experience, education, honors, volunteering, military background, technical skills, and employment gaps for resumes; defendant gender, ethnicity, age, prior convictions, juvenile record, and charge degree for recidivism).

Interpreting Uncertainty in Posterior Distributions

As part of our validation, we present posterior distributions $P(\theta|\{y_D\})$ for each guardrail's activation parameter θ_i. The top density plots (e.g., Fig. 2) show the posterior mean and 94% Highest Density Interval (HDI), indicating the range where the most credible 94% of posterior θ values lie, reflecting uncertainty in the likelihood of guardrail activation.

We sampled joint vectors $\hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_N]$ from the posterior $P(\theta|\{y_\mathcal{D}\})$, using $\hat{\theta}_i$ as the success probability in a Bernoulli trial to decide guardrail activation. Repeating this 1,000 times yielded histograms of activation outcomes (e.g., bottom plots in Fig. 2), revealing how often each guardrail is selected vs. rejected and which combinations best align model outputs with expert expectations.

Baseline Strategies for Guardrail Selection

We compared our guardrail selection methods against two baselines (Table 2 and Table 4). The first uses the LLM without guardrails as a reference to assess alignment and predictive performance against expert-labeled ground truths. The second, "fixed best guardrail combination", exhaustively evaluates all 2^N combinations, selects the one with the highest F1 score, and applies it uniformly to all records. While this provides an upper performance bound, it is computationally expensive and ignores instance-level variation. Our method outperforms it by sampling θ from the posterior per record and activating guardrails probabilistically.

Evaluation Metric

The evaluation metric for our method targets task-specific alignment—how well selected guardrails align LLM outputs $\hat{y}$ with expert knowledge. Although existing fairness metrics (Hardt, Price, and Srebro 2016) assess systemic bias, our validation datasets $\mathcal{D}$ are demographically balanced and include expert-informed ground truths $y_\mathcal{D}$ labeled by experts or validated real-world outcomes, reflecting fairness in the alignment objective. We therefore use the F1 score (Rijsbergen 1979) (harmonic mean of precision and recall) to assess the effectiveness of guardrail selection for user-LLM alignment. The first author computed all F1 scores, and all three authors verified them through iterative discussions.

After learning posterior distributions over θ, the surrogate model was no longer needed for validation. Instead, we used posterior θ samples to select a guardrail configuration per record, enabling direct LLM queries without evaluating all combinations. We computed the final F1 scores by comparing all $\{\hat{y}\}$ against $\{y_\mathcal{D}\}$ (Table 2 and Table 4).

Equal Opportunity in Hiring

Predicting resume quality improves candidates' chances of securing interviews and jobs. However, algorithmic assessments risk bias from demographics (Arzaghi, Carichon, and Farnadi 2025), pronouns (Ghosh and Caliskan 2023), or hiring context (Li et al. 2021). Our pre-experiments showed LLMs predicted gender and ethnicity from names with 0.9 and 0.98 F1 scores, respectively. We used this task to evaluate how moderation guardrails align LLM-predicted resume quality with expert assessments, promoting best hiring practices and equal opportunity regardless of gender or ethnicity.

Resume Quality Prediction Dataset: We used a publicly available dataset (Bertrand and Mullainathan 2004) $\mathcal{D}$, with resume quality ("high" or "low") labeled by professional evaluators. This dataset contains an equal distribution of high- and low-quality resumes across genders and ethnicities. It covers Chicago and Boston job postings from 2001–2002, with roles distributed across business (27%), trade (21%), health/education/social services (15%), finance (8%), manufacturing (8%), transport (3%), and other (18%).

Resume Quality Prediction Guardrails: We implemented guardrails G^{RQ} (Table 1) based on expert hypotheses (G_1^{RQ} and G_2^{RQ}) and empirical findings (G_3^{RQ}) from existing work (Bertrand and Mullainathan 2004) to align the LLM predictions with key resume assessment criteria. We also added guardrails (G_4^{RQ}, G_5^{RQ}, and G_6^{RQ}) that explicitly instruct the LLM to ignore names and avoid inferring demographics during predictions. We also included a distractor guardrail (G_7^{RQ}) that contradicts empirical findings (G_3^{RQ}), promoting features known to indicate low resume quality. While guardrail engineers may not deploy such distractor guardrails in practice since they are easily identifiable, we still intentionally included them to test whether our method can identify and reject them, simulating guardrail design failures and serving as a robustness check.

ID	Guardrails to moderate resume quality predictions	Guardrail type
G_1^{RQ}	When judging a candidate's qualifications, use criteria such as labor market experience, career profile, the existence of gaps or holes in employment, and skills listed.	*Expert hypothesis* (general features)
G_2^{RQ}	High-quality resumes have the following: summer or while-at-school employment experience, volunteering experience, extra computer skills, certification degrees, foreign language skills, honors, or some military experience.	*Expert hypothesis* (specific criteria or feature values)
G_3^{RQ}	Higher-quality applicants have, on average, a little more labor market experience and fewer holes in their employment history; they are also more likely to have an e-mail address, have completed a certification degree, possess foreign language skills, or have been awarded some honors.	*Empirical findings* (specific criteria or feature values)
G_4^{RQ}	When judging an applicant's resume quality, do not consider the applicant's name, even if it is mentioned in the resume.	*Counter bias* (applicant's name)
G_5^{RQ}	When judging an applicant's resume quality, do not infer the applicant's gender from any features in the resume, and do not consider it even if gender is mentioned in the resume.	*Counter bias* (applicant's gender)
G_6^{RQ}	When judging an applicant's resume quality, do not infer the applicant's ethnicity from any features in the resume, and do not consider it even if ethnicity is mentioned in the resume.	*Counter bias* (applicant's ethnicity)
G_7^{RQ}	Higher-quality applicants have, on average, less labor market experience and more holes in their employment history; they are also less likely to have an e-mail address, have not completed a certification degree, do not possess foreign language skills, or have not been awarded some honors.	*Distractor* for G_3^{RQ} (opposite criteria or feature values)

Table 1: Guardrails applied to moderate LLM outputs for resume quality prediction (Bertrand and Mullainathan 2004).

Resume Quality Prediction Using Surrogate Model: Instead of querying the LLM directly $623{,}360$ times (i.e., once for each of the $4{,}870$ unique resumes in our dataset and $2^7 = 128$ unique guardrail activation indicator vectors $\mathbf{g}$ for our 7 guardrails), we replicated LLM-predicted resume quality using a surrogate DNN model. Our surrogate DNN consisted of three 64-unit ReLU layers with batch normalization, dropout, and a sigmoid output trained via Adam. We generated 102,400 DNN training samples by selecting 800 resumes for each of the possible 128 guardrail activation indicator vectors. To ensure balanced coverage across guardrails, we used stratified sampling based on balanced information acquisition principles (Krause and Guestrin 2005). To get binary LLM resume prediction labels (i.e., "ground truth") for each of the samples, we queried the LLM using a system prompt with the sample's corresponding $\mathbf{g}$ and a user prompt from the raw resume $x_{\mathcal{D}}$. We then created the surrogate DNN dataset, with each record combining the corresponding resume encoding x_{encode}, guardrail activation indicator vector $\mathbf{g}$, and binary LLM prediction representing one of the 102,400 samples. Using a 60%/10%/30% stratified train/validation/test split, the grid-search-tuned model achieved a 91% F1 score on the held-out test set.

Selecting Effective Guardrails: We first investigated the combined influence (Fig. 2) of all resume quality guardrails from Table 1. Results show that G_3^{RQ} and G_6^{RQ} yield the strongest alignment, highlighting the effectiveness of pairing empirical evidence-focused guardrails with demographic bias mitigation. In contrast, adding G_7^{RQ} reduced alignment, likely due to conflicting emphasis with G_3^{RQ}. We then conducted an ablation analysis to explore guardrail interactions.

Identifying Complementary Guardrails: We consider G_1^{RQ} and G_2^{RQ} together since both stem from expert hypothesis (Fig. 3). G_1^{RQ} identifies general features (e.g., work experience, skills) for resume quality but not their specific values, slightly helping alignment. G_2^{RQ} specifies feature values associated with high-quality resumes, yielding stronger alignment. Together, G_1^{RQ} ensures broad coverage, while G_2^{RQ} sharpens focus on key indicators, creating a balanced mix.

Identifying Individual Optimal Contributions: Next, we compared expert hypothesis-based guardrails (G_1^{RQ} and G_2^{RQ}) with the empirical evidence-based guardrail G_3^{RQ} (Fig. 4). Even though G_1^{RQ}, G_2^{RQ}, and G_3^{RQ} individually perform well, combining multiple overlapping guardrails can introduce redundancy and lead to diminishing returns. Results show that G_3^{RQ}, which directly encodes high-quality resume features based on evidence, is the most effective. However, adding G_1^{RQ} and G_2^{RQ} introduces redundancy.

Rejecting Distractor Guardrails: Since G_3^{RQ} showed the strongest alignment, we tested it against a distractor guardrail G_7^{RQ} (Fig. 5). G_7^{RQ} contradicts hiring criteria, claiming high-quality candidates have negative features (e.g., employment gaps). Our method selected G_3^{RQ} (empirical evidence-based) and rejected G_7^{RQ} (distractor).

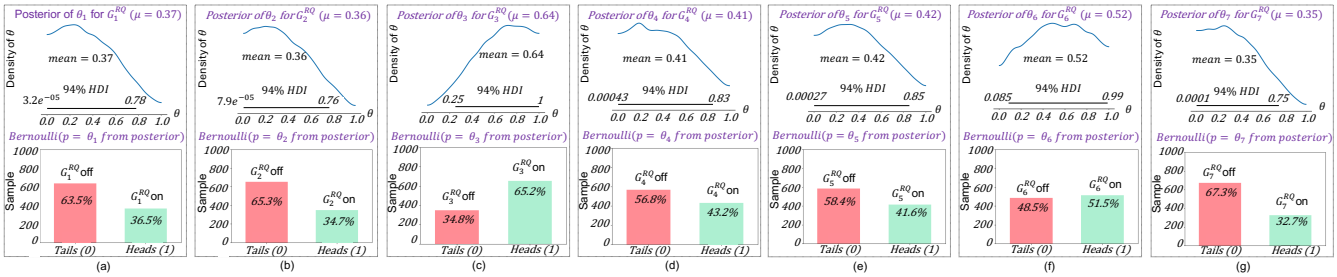

Figure 2: Posterior distributions $P(\theta|\{y_D\})$ and 1,000 Bernoulli toss outcomes using sampled θ values as success probabilities ($p = \theta$) to decide activation of resume quality guardrails: (a) G_1^{RQ}, (b) G_2^{RQ}, (c) G_3^{RQ}, (d) G_4^{RQ}, (e) G_5^{RQ}, (f) G_6^{RQ}, and (g) G_7^{RQ}.

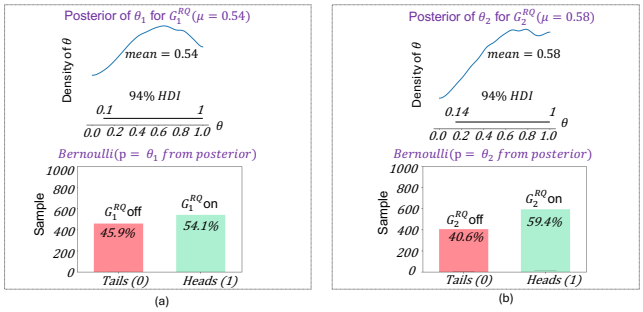

Figure 3: Our method effectively selected complementary resume quality guardrails: a) G_1^{RQ} and b) G_2^{RQ}.

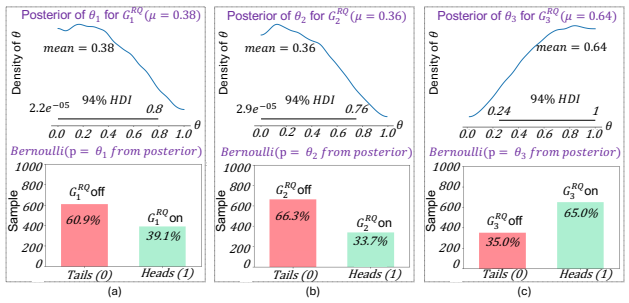

Figure 4: When combining multiple resume quality guardrails, our method rejected both a) G_1^{RQ} and b) G_2^{RQ} due to overlapping features with c) G_3^{RQ}.

Influence of Counter-Bias Guardrails: We compared counter-bias guardrails for name (G_4^{RQ}), gender (G_5^{RQ}), and ethnicity (G_6^{RQ}) (Fig. 6). G_4^{RQ} had the strongest impact, as names are key demographic proxies, whereas the effects of G_5^{RQ} and G_6^{RQ} were less pronounced—either because gender and ethnicity are not considered part of resumes, even if disclosed, or the LLM already incorporates in-built bias-mitigation guardrails. If names are removed using an anonymization tool, our method would mark the name-based guardrail indifferent, but the model could still infer gender and ethnicity from proxy features like education or work history, requiring stronger counter-bias guardrails.

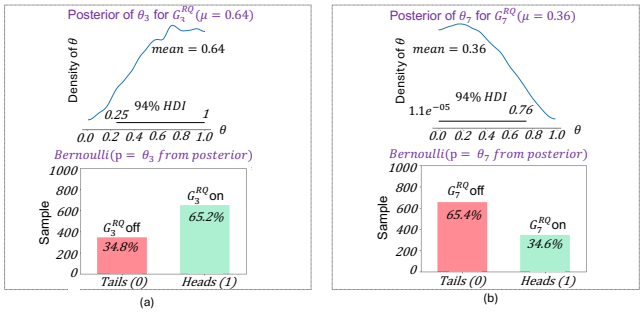

Figure 5: Our method selected the best resume quality guardrail a) G_3^{RQ} and rejected its opposite distractor b) G_7^{RQ}.

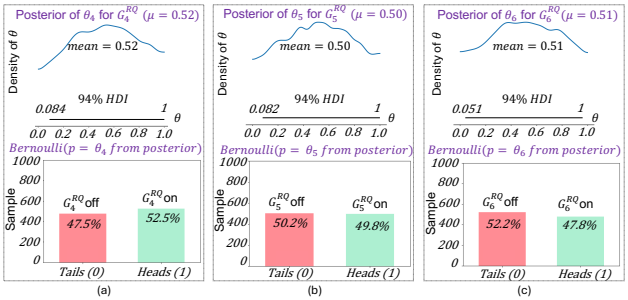

Figure 6: Counter-bias resume quality guardrails: a) G_4^{RQ}, b) G_5^{RQ}, and c) G_6^{RQ} showed limited impact likely due to LLM's training-stage bias mitigation of name, gender, or ethnicity.

LLM Predictive Performance: Our method selects guardrails that improve resume quality predictions (F1 score: 0.71) and align with expert labels $\{y_{\mathcal{D}}\}$ (Table 2).

Fairness and Criminal Justice

Predicting recidivism (i.e., a defendant's likelihood of reoffending) is critical in risk assessment, sentencing, and parole decisions. We selected this task to assess how moderation guardrails align LLM-predicted recidivism with actual reoffense outcomes while promoting fairness in criminal justice.

Recidivism Prediction Dataset: To create our alignment dataset $\mathcal{D}$, we used a publicly available COMPAS

Guardrail selection method for resume quality	F1 score
Model with no guardrail	0.65
Fixed best guardrail combination	0.68
Bayesian MCMC (mean of $P(\theta\|\{y_D\})$ as cut-off)†	0.68
Bayesian MCMC (sampled θ as cut-off)†	0.69
Bayesian MCMC (Bernoulli toss with sampled θ)†	0.71

Table 2: LLM performance on resume quality prediction using our proposed† and baseline guardrail selection methods.

dataset (Mallari et al. 2020) of Broward County, Florida criminal records from 2013–2014 (35.3% misdemeanors, 64.7% felonies). It reports whether defendants re-offended within two years of initial screenings. We removed records with missing outcomes (reducing 7,214 to 6,172) and applied up-sampling to ensure balanced recidivism outcomes across gender and ethnicity (resulting in 7,310 records).

Recidivism Prediction Guardrails: We construct guardrails G^{RC} informed by COMPAS guide (Northpointe 2015) and existing literature findings (Dressel and Farid 2018; Mallari et al. 2020) on recidivism prediction (Table 3). These guardrails emphasize established risk factors, including prior criminal history (G_1^{RC}), degree of involvement (G_2^{RC}), dynamic personality traits (G_3^{RC}), and violent behavior patterns (G_4^{RC}). After consulting two external researchers, we added counter-bias guardrails (G_5^{RC}, G_6^{RC}, and G_7^{RC}) to prevent the LLM from using the defendant's name, gender, or ethnicity. Although such demographic factors are accessible to judges, they are not directly predictive of recidivism risk. Since we already demonstrated in the hiring task that our method can identify and reject distractor guardrails, we did not repeat that analysis for this dataset.

ID	Guardrails to moderate recidivism prediction	Guardrail Type
G_1^{RC}	Well-known recidivism risk factors involve criminal associates, criminal personality, and drug involvement, according to various prior studies of recidivism.	*Empirical findings* (general factors)
G_2^{RC}	The degree of criminal involvement has consistently emerged as a major risk factor for predicting ongoing criminal behavior. It is the most important of the major risk factors emerging in various meta-analysis studies. Early juvenile delinquency involvement is also linked to ongoing criminal behavior, according to prior studies.	*Empirical findings* (degree of criminal involvement)
G_3^{RC}	Criminal personality (e.g., impulsivity, no guilt, selfishness/-narcissism, a tendency to dominate others, risk-taking, and a history of violent behavior or aggression) is the second most important dynamic factor in predicting recidivism.	*Empirical findings* (characteristics of criminal personality)
G_4^{RC}	A history of violent behavior has been demonstrated to be one of the most powerful predictors of future violence. The likelihood of future violence appears to steadily increase with each instance of a prior violent incident. Each prior arrest for violent behavior increases the likelihood of further violence. Similarly, a history of juvenile violence has been found to be a predictor of adult violence.	*Empirical findings* (prior history of juvenile violence)
G_5^{RC}	When predicting recidivism, do not consider the defendant's name, even if it is mentioned in the document.	*Counter bias* (defendant's name)
G_6^{RC}	When predicting recidivism, do not infer the defendant's gender from any features in the resume, and do not consider the defendant's gender, even if mentioned in the document.	*Counter bias* (defendant's gender)
G_7^{RC}	When predicting recidivism, do not infer the defendant's ethnicity and race from any features in the resume, and do not consider the defendant's ethnicity and race, even if these are mentioned in the document.	*Counter bias* (defendant's ethnicity and race)

Table 3: Guardrails applied to moderate LLM outputs for recidivism prediction (Northpointe 2015; Mallari et al. 2020).

Recidivism Prediction Using Surrogate Model: Instead of querying the LLM directly $935{,}680$ times (i.e., once for each of the $7{,}310$ unique defendants in our dataset and $2^7 = 128$ unique guardrail activation indicator vectors $\mathbf{g}$ for our 7 guardrails), we replicated LLM-predicted recidivism using a surrogate DNN model. The DNN consisted of five 64-unit ReLU layers with batch normalization, dropout, and a sigmoid output trained via Adam. We generated $153{,}600$ DNN training samples by selecting $1{,}200$ defendant records for each of the possible 128 guardrail activation indicator vectors. To ensure balanced coverage across guardrails, we used stratified sampling based on balanced information acquisition principles (Krause and Guestrin 2005). To get binary LLM resume prediction labels (i.e., "ground truth") for each of the samples, we queried the LLM using a system prompt with the sample's guardrail activation vector and a user prompt from the raw defendant record $x_{\mathcal{D}}$. We then created the surrogate DNN dataset, with each record combining the corresponding defendant record encoding x_{encode}, guardrail activation indicator vector $\mathbf{g}$, and binary LLM prediction for one of the $153{,}600$ samples. Using a 60%/10%/30% stratified train/validation/test split, the grid-search-tuned model achieved a 93.4% F1 score on the held-out test set.

Selecting Effective Guardrails: We applied our method to all recidivism guardrails from Table 3 to assess their combined influence (Fig. 7). The best alignment with $\mathcal{D}$ was achieved using G_1^{RC}, G_2^{RC}, and G_4^{RC} that capture key predictive signals like prior criminal history, criminal involvement, and violent behavior. In contrast, adding G_5^{RC}, G_6^{RC}, and G_7^{RC}, which were supposed to reduce biases, slightly reduced alignment, likely due to stricter exclusion criteria. To further analyze guardrail interactions, we conducted ablation experiments with selected subsets of guardrails.

Identifying Complementary Guardrails: The COMPAS practitioner's guide-based guardrails (Northpointe 2015) collectively improved model alignment in recidivism prediction (Fig. 8). Our method selected G_1^{RC}, G_2^{RC}, and G_4^{RC} as the most effective combination, rejecting G_3^{RC}, likely due to redundant overlapping features with G_1^{RC} and G_4^{RC}.

Identifying Counter-bias Guardrails Our results show that counter-bias guardrails (G_5^{RC}, G_6^{RC}, and G_7^{RC}) did not improve alignment and sometimes performed worse than using no guardrails (Fig. 9). These guardrails were designed to suppress the influence of demographic attributes like defendants' names, genders, and ethnicity in recidivism prediction. However, doing so may have removed useful context, as such attributes can inform nuanced recidivism decisions. This suggests that completely removing demographic information may misalign with expert decisions, such as those made by criminal justice decision-makers informed by risk assessment tools (Larson et al. 2016; Northpointe 2015), who may legitimately consider such context. This contrasts with resume screening, where demographic features are more likely to introduce bias than reduce it.

LLM Predictive Performance: Our method selects guardrails that improve recidivism predictions (F1 score: 0.66) and align with expert labels $\{y_{\mathcal{D}}\}$ (Table 4).

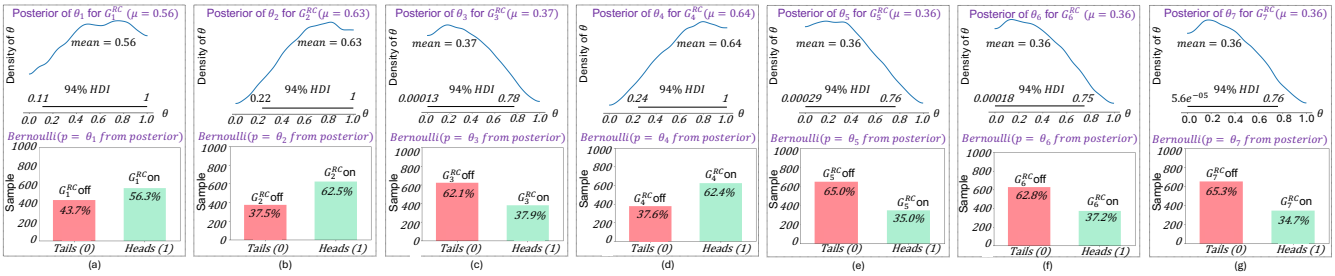

Figure 7: Posterior distributions $P(\theta|\{y_D\})$ and 1,000 Bernoulli toss outcomes using sampled θ values as success probabilities ($p = \theta$) to decide activation of recidivism guardrails: (a) G_1^{RC}, (b) G_2^{RC}, (c) G_3^{RC}, (d) G_4^{RC}, (e) G_5^{RC}, (f) G_6^{RC}, and (g) G_7^{RC}.

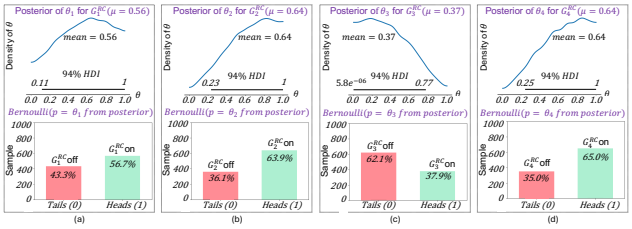

Figure 8: Our method selects: a) G_1^{RC}, b) G_2^{RC}, and d) G_4^{RC}, and rejects c) G_3^{RC} likely due to overlaps with G_1^{RC} and G_4^{RC}.

Guardrail selection method for recidivism	F1 score
Model with no guardrail	0.62
Fixed best guardrail combination	0.63
Bayesian MCMC (mean of $P(\theta\|\{y_D\})$ as cut-off)†	0.63
Bayesian MCMC (sampled θ as cut-off)†	0.65
Bayesian MCMC (Bernoulli toss with sampled θ)†	0.66

Table 4: LLM performance comparing our proposed† and baseline guardrail selection methods to predict recidivism.

Discussion and Conclusion

In this work, we presented a probabilistic framework to evaluate and select output moderation guardrails for aligning LLM outputs with expert user expectations. By leveraging Bayesian inference, we estimated guardrail activation probabilities and interpreted their contributions to model alignment both independently and in combination with others.

Our validation showed our method's ability to identify effective guardrails, reject distractors, and address biases in resume quality and recidivism prediction tasks. Our results showed that our approach improves model alignment with empirical data while enabling a nuanced understanding and interpretation of how guardrails influence LLM outputs.

We found that it is not always true that guardrails that are individually better will also work the same when combined with other guardrails. Rather, sometimes, they may cancel the effect of another guardrail. Also, if the base LLM model has already aligned to expert expectations during its training, it may happen that a similarly constructed user guardrail does not have much influence on changing the LLM outputs.

We demonstrated our method using moderation guardrails to ensure equal opportunity in resume quality and fairness

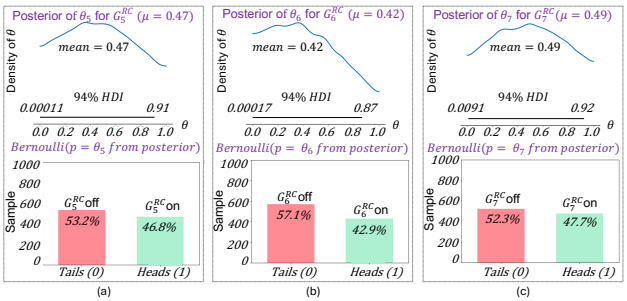

Figure 9: Our method rejects counter-bias recidivism guardrails: a) G_5^{RC}, b) G_6^{RC}, c) G_7^{RC}, suggesting that fully ignoring demographic features may reduce alignment.

in recidivism prediction. However, future work should study the applicability of our method to other guardrails (e.g., preventing jailbreaking, hallucination) with different alignment goals (e.g., security, safety, and adversarial robustness).

We conducted quantitative evaluations to select effective guardrails. Future work could engage domain experts for qualitative studies and integrate explanation tools to improve local (per-resume) and cohort-level (across-demographics) interpretability of guardrail effects toward alignment.

Our method could target domain experts, AI ethicists, policymakers, and industry practitioners, depending on the alignment task and evaluation criteria. Since alignment goals may differ or even conflict across stakeholders, our method emphasizes defining the alignment objective and target audience first. If the alignment task involves regulatory compliance (e.g., U.S. federal AI governance (U.S. Congress 2024), the EU AI Act (European Parliament and Council 2024)), our method can systematically extend to evaluate guardrails designed to meet such policies, provided the alignment dataset incorporates expert input from policymakers. This could support auditability, traceability, and non-discrimination in high-risk applications.

Future work should explore scaling our approach to larger datasets, complex tasks, and adaptive guardrail configurations that respond to evolving user needs. These efforts will further improve "user-LLM" alignment and enhance the practical and ethical deployment of LLM-based decision-support systems in high-stakes real-world applications.

Ethical Statement

This research aligns with AIES's commitment to ethical AI development by ensuring responsible dataset usage and mitigating potential societal harms. We evaluate and select guardrails to improve "user-LLM" alignment in high-stakes decision-making, particularly in domains like hiring and criminal justice. Our method relies on publicly available datasets, following ethical guidelines. While our approach aims to reduce biases, we acknowledge that no method can fully eliminate unintended harms. Future work will explore additional security-focused guardrails, alternative alignment strategies, and broader societal implications to enhance the responsible deployment of LLMs in decision-making.

Broader Impact

While our method aims to improve alignment in LLM-driven decision-making with standards set by subject matter authority experts, we need to be aware of unintended risks if misapplied. There is a risk that probabilistically selected guardrails may be treated as universally valid, potentially reinforcing dominant norms and marginalizing alternative viewpoints. Institutions and organizations could use alignment claims to justify biased practices if alignment goals are narrowly defined or poorly contextualized. We urge practitioners to apply this method with transparency, domain expertise, inclusive, and broader stakeholder engagement.

Positionality Statement

As researchers in human-centered AI, we acknowledge our positionality in shaping the design, evaluation, and interpretation of this work. Our interdisciplinary background spans computer science, human-computer interaction (HCI), computational modeling, and uncertainty quantification (UQ), informing our approach to aligning LLM behavior with expert expectations in high-stakes domains. While we are not domain experts in hiring or criminal justice, we grounded our work in established literature, followed ethical guidelines, got researchers' feedback, and consulted relevant empirical findings to design alignment datasets and guardrails. We recognize that value judgments and sociotechnical context play a role in defining "expert user expectations," and we aim to make our methods transparent and adaptable to diverse perspectives. Our goal is to support AI practitioners in critically evaluating and selecting moderation guardrails that reflect AI fairness and accountability in decision-making.

Declaration on the Use of AI

For our experiments, we used our institutional LLM, U-M GPT, provided by the University of Michigan, with API access to the GPT-4o model to evaluate and select effective guardrails that better align LLM behaviors with expert user expectations. We *did not* use generative AI tools to write, revise, or generate this paper. All text, analysis, and figures were produced by the authors.

Acknowledgements

We thank members of the Computational HCI (CompHCI) Lab at the University of Michigan for informal brainstorming. We thank Divya Ramesh and Nel Escher for informing the dataset selection and alignment criteria, and Tsedeniya Amare for proofreading the paper.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Arzaghi, M.; Carichon, F.; and Farnadi, G. 2025. *Understanding Intrinsic Socioeconomic Biases in Large Language Models*, 49–60. AAAI Press.

Arzberger, A.; Buijsman, S.; Lupetti, M. L.; Bozzon, A.; and Yang, J. 2025. Nothing Comes without Its World - Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. In *Proc. of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, 61–73. AAAI Press.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The Moral Machine experiment. *Nature*, 563: 59–64.

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 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA. ISBN 9781450383097.

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4): 991–1013.

Boggust, A.; Hoover, B.; Satyanarayan, A.; and Strobelt, H. 2022. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *Proceedings of the 2022 ACM CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA. ISBN 9781450391573.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Chiang, C.-W.; Lu, Z.; Li, Z.; and Yin, M. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA. ISBN 9781450394215.

Chiang, C.-W.; Lu, Z.; Li, Z.; and Yin, M. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th ACM International Conference on Intelligent User Interfaces*, IUI '24. NY. ISBN 9798400705083.

Chivukula, S. S.; Gray, C.; Li, Z.; Pivonka, A. C.; and Chen, J. 2024. Surveying a Landscape of Ethics-Focused Design Methods. *ACM J. Responsib. Comput.*, 1(3).

Chu, Z.; Wang, Y.; Li, L.; Wang, Z.; Qin, Z.; and Ren, K. 2024. A Causal Explainable Guardrails for Large Language Models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24. NY. ISBN 9798400706363.

Cornacchia, G.; Zizzo, G.; Fraser, K.; Hameed, M. Z.; Rawat, A.; and Purcell, M. 2025. *MoJE: Mixture of Jailbreak Experts, Naive Tabular Classifiers as Guard for Prompt Attacks*, 304–315. AAAI Press.

Deshpande, A.; and Sharp, H. 2022. Responsible AI Systems: Who are the Stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 227–236. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Dev, J.; Akhuseyinoglu, N.; Kayas, G.; Rashidi, B.; and Garg, V. 2024. Building Guardrails in AI Systems with Threat Modeling. *Digit. Gov.: Res. Pract.* Just Accepted.

Dong, Y.; Mu, R.; Jin, G.; Qi, Y.; Hu, J.; Zhao, X.; Meng, J.; Ruan, W.; and Huang, X. 2024. Position: building guardrails for large language models requires systematic design. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.*, 4(1): eaao5580.

Dutta, A.; Khorramrouz, A.; Dutta, S.; and KhudaBukhsh, A. R. 2024. Down the toxicity rabbit hole: a framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24. ISBN 978-1-956792-04-1.

Estrada, D. 2018. Value Alignment, Fair Play, and the Rights of Service Robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 102–107. New York, NY, USA: Association for Computing Machinery. ISBN 9781450360128.

European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 12.7.2024, p. 1–254, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.

Fang, C.; Qin, C.; Zhang, Q.; Yao, K.; Zhang, J.; Zhu, H.; Zhuang, F.; and Xiong, H. 2023. RecruitPro: A Pretrained Language Model with Skill-Aware Prompt Learning for Intelligent Recruitment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23. New York, NY, USA. ISBN 9798400701030.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3).

Ghosh, S.; and Caliskan, A. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 901–912. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Gombolay, M. 2024. Human-robot alignment through interactivity and interpretability: don't assume a "spherical human". In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24. ISBN 978-1-956792-04-1.

Hadfield-Menell, D.; Andrus, M.; and Hadfield, G. 2019. Legible Normativity for AI Alignment: The Value of Silly Rules. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 115–121. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Hadfield-Menell, D.; and Hadfield, G. K. 2019. Incomplete Contracting and AI Alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 417–422. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3323–3331. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Hrytsyna, A.; and Alves, R. 2024. From Representation to Response: Assessing the Alignment of Large Language Models with Human Judgment Patterns. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; and Khabsa, M. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. *ArXiv*, abs/2312.06674.

Iqbal, U.; Kohno, T.; and Roesner, F. 2025. *LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins*, 611–623. AAAI Press.

Kato, M.; Imaizumi, M.; and Minami, K. 2023. Unified Perspective on Probability Divergence via the Density-Ratio Likelihood: Bridging KL-Divergence and Integral Probability Metrics. In Ruiz, F.; Dy, J.; and van de Meent, J.-W., eds., *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 5271–5298. PMLR.

Krause, A.; and Guestrin, C. 2005. Near-Optimal Nonmyopic Value of Information in Graphical Models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, 324–331. Arlington, Virginia, USA: AUAI Press. ISBN 0974903914.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Lai, V.; Chen, C.; Smith-Renner, A.; Liao, Q. V.; and Tan, C. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23. NY, USA. ISBN 9798400701924.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. Published May 9, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Li, L.; Lassiter, T.; Oh, J.; and Lee, M. K. 2021. Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 166–176. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Li, T.; Khashabi, D.; Khot, T.; Sabharwal, A.; and Srikumar, V. 2020. UNQOVERing Stereotyping Biases via Underspecified Questions. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3475–3489. Online: Association for Computational Linguistics.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634.

MacKay, D. J. C. 2002. *Information Theory, Inference & Learning Algorithms*. USA: Cambridge University Press. ISBN 0521642981.

Mahomed, Y.; Crawford, C. M.; Gautam, S.; Friedler, S. A.; and Metaxa, D. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. NY, USA. ISBN 9798400704505.

Mallari, K.; Inkpen, K.; Johns, P.; Tan, S.; Ramesh, D.; and Kamar, E. 2020. Do I Look Like a Criminal? Examining how Race Presentation Impacts Human Judgement of Recidivism. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–13. New York, NY, USA. ISBN 9781450367080.

Norhashim, H.; and Hahn, J. 2024. Measuring Human-AI Value Alignment in Large Language Models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1063–1073.

Norhashim, H.; and Hahn, J. 2025. *Measuring Human-AI Value Alignment in Large Language Models*, 1063–1073. AAAI Press.

Northpointe, I. 2015. Practitioner's Guide to compas Core.

Pangakis, N.; and Wolken, S. 2024. Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels. In Card, D.; Field, A.; Hovy, D.; and Keith, K., eds., *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, 113–131. Mexico City, Mexico: Association for Computational Linguistics.

Prabhudesai, S.; Yang, L.; Asthana, S.; Huan, X.; Liao, Q. V.; and Banovic, N. 2023. Understanding Uncertainty: How Lay Decision-Makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th ACM International Conference on Intelligent User Interfaces*, IUI '23, 379–396. New York, NY, USA. ISBN 9798400701061.

Rastogi, C.; Tulio Ribeiro, M.; King, N.; Nori, H.; and Amershi, S. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 913–926. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Rebedea, T.; Dinu, R.; Sreedhar, M. N.; Parisien, C.; and Cohen, J. 2023. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In Feng, Y.; and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 431–445. Singapore.

Rijsbergen, C. J. V. 1979. *Information Retrieval*. USA: Butterworth-Heinemann, 2nd edition. ISBN 0408709294.

Robinson, P. 2023. Action Guidance and AI Alignment. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 387–395. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Shankar, S.; Zamfirescu-Pereira, J.; Hartmann, B.; Parameswaran, A.; and Arawjo, I. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In *Proceedings of the 37th ACM Symposium on User Interface Software and Technology*, UIST '24. NY. ISBN 9798400706288.

Shneiderman, B. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Trans. Interact. Intell. Syst.*, 10(4).

Shree, S.; Khadka, K.; Lei, Y.; Kacker, R. N.; and Kuhn, D. R. 2024. Constructing Surrogate Models in Machine Learning Using Combinatorial Testing and Active Learning. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ASE '24, 1645–1654. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712487.

Terry, M.; Kulkarni, C.; Wattenberg, M.; Dixon, L.; and Morris, M. R. 2024. Interactive AI Alignment: Specification, Process, and Evaluation Alignment. arXiv:2311.00710.

U.S. Congress. 2024. Federal A.I. Governance and Transparency Act of 2024, H.R. 7532, 118th Congress (2023–2024), amending Title 44 of the United States Code. Public Law, March 2024. Text available at: https://www.congress.gov/bill/118th-congress/house-bill/7532. Adds Subchapter IV—Artificial Intelligence System Governance to Title 44, U.S. Code.

Varshney, K. R. 2025. *Decolonial AI Alignment: Openness, Viśeundefineda-Dharma, and Including Excluded Knowledges*, 1467–1481. AAAI Press.

Wang, J.; Hu, H.; Wang, Z.; Yan, S.; Sheng, Y.; and He, D. 2024a. Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. In *Proceedings of the 2024 ACM CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA. ISBN 9798400703300.

Wang, X.; Duan, S.; Yi, X.; Yao, J.; Zhou, S.; Wei, Z.; Zhang, P.; Xu, D.; Sun, M.; and Xie, X. 2024b. On the

essence and prospect: an investigation of alignment approaches for big models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24. ISBN 978-1-956792-04-1.

Wang, Z.; Yang, F.; Wang, L.; Zhao, P.; Wang, H.; Chen, L.; Lin, Q.; and Wong, K.-F. 2024c. SELF-GUARD: Empower the LLM to Safeguard Itself. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1648–1668. Mexico.

Yang, Z.; Wu, Y.; Wen, R.; Backes, M.; and Zhang, Y. 2025. Peering Behind the Shield: Guardrail Identification in Large Language Models. arXiv:2502.01241.

Yuan, Z.; Xiong, Z.; Zeng, Y.; Yu, N.; Jia, R.; Song, D.; and Li, B. 2024. RigorLLM: resilient guardrails for large language models against undesired content. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhang, Y.; Chen, Z.; Fang, Y.; Lu, Y.; Li, F.; Zhang, W.; and Chen, H. 2024. Knowledgeable Preference Alignment for LLMs in Domain-specific Question Answering. arXiv:2311.06503.

Zhao, H.; Andriushchenko, M.; Croce, F.; and Flammarion, N. 2024. Long is more for alignment: a simple but tough-to-beat baseline for instruction fine-tuning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Zheng, M.; Pei, J.; Logeswaran, L.; Lee, M.; and Jurgens, D. 2024. When "A Helpful Assistant" Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15126–15154. Miami, Florida, USA: Association for Computational Linguistics.

Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2024. Beyond Preferences in AI Alignment. *Philosophical Studies*.

Zhou, Y.; Liu, X.; Ning, C.; and Wu, J. 2024. MultifacetEval: Multifaceted Evaluation to Probe LLMs in Mastering Medical Knowledge. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6669–6677.