CALIBRATING LLM EXPLANATIONS TO HUMAN DECISIONS WITH REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Explanations of LLM outputs are considered an important way to make model behavior legible, helping human users make good decisions on downstream tasks. However, default training approaches for improving explanations target model task performance or human preference rather than human decision quality. To bridge this gap, we introduce **HD-RL**, a reinforcement learning framework for optimizing LLM explanations for human decision making. Our approach calibrates a human decision simulator on observed human decisions, applying a proxy reward model to optimize the LLM to generate explanations that help people to make good decisions. We apply this framework to AI-generated text detection. In our initial data collection phase, we observed that while LLMs possess strong autonomous detection capabilities, the explanations they generate for users yield near-random human accuracy (51.3%). By applying the HD-RL framework to the open-source Qwen 3.6 model, we generate optimized explanations that improve human performance, achieving a GLMM-estimated human accuracy of 63.7% in a randomized between-subjects experiment with repeated measures and outperforming RLVR, RLAIF, and other baselines by over 8%. Full code and data are available at `https://anonymous.4open.science/r/Human_Decision_RL-2471/`.

1 INTRODUCTION

As large language models (LLMs) become increasingly powerful, people rely on them more heavily for complex, everyday decision-making tasks (Kim et al., 2025; Duenas & Ruiz, 2024). When collaborating with an LLM, users frequently look to the model's explanations or reasoning traces to help them assess how it might have arrived at a particular conclusion. Even when explanations are not fully faithful to the model's reasoning, they play an important role in people's decisions by mediating trust (Ribeiro et al., 2016).

However, default approaches to improving LLM explanations, such as CoT prompting (Wei et al., 2022) or standard preference alignment (Christiano et al., 2017; Rafailov et al., 2023; Turpin et al., 2023; Lanham et al., 2023; Li et al., 2026), target objectives that are only indirectly related to human users' decision quality, such as the model's autonomous task performance or ability to produce explanations that appeal to human preferences. These approaches can result in distinct failure modes for human use of explanations. For example, optimizing an LLM to maximize its own autonomous task accuracy (e.g., via verifiable-reward RL) only ensures the reasoning trace is logically legible to the model itself. It does not guarantee decision utility to an end-user. Similarly, optimizing for subjective human preference conflates preference with decision utility, because people frequently favor intuitively plausible explanations over practically useful ones Palta et al. (2026); Oliveira et al. (2025). Recent literature demonstrates that plausible-sounding LLM explanations often degrade human judgment rather than enhance it (Wang et al., 2026; Cheung et al., 2025; Wang et al., 2025).

Statistical decision theory (Berger, 1987) provides a natural framework for formulating the goal of explanations intended as human decision support. A decision-maker observes information about an uncertain state of the world, including an explanation, and chooses an action under a utility function. The quality of the explanation is evaluated by whether it helps the decision-maker select a good action under the utility. Hence, to train explanations for good downstream decision use, methods should target the distribution over actions of human decision-makers who use that explanation.

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

We contribute a decision-theoretic framework for explanation post-training that distills this intuition to a measurable target. Our approach first trains a decision simulator on a large-scale human decision dataset to predict the probability distribution of human decisions conditioned on a given text and explanation. We then use this simulator as a proxy reward model to optimize an LLM policy via reinforcement learning. To construct the reward signal, we evaluate the simulator's predicted probability distribution over the human's actions against the objective ground-truth label, generating a logarithmic surrogate reward. By optimizing the policy to maximize this reward, the LLM is explicitly trained to generate explanations that help people to make good decisions.

We demonstrate the HD-RL framework on an AI text detection task. We collect and release a large-scale dataset comprising 21K real-world human decisions on 2712 unique texts, where half are organically written by human. In a randomized between-subjects experiment with repeated measures, Generalized Linear Mixed Model (GLMM) results show that explanations generated via HD-RL achieve an estimated human detection accuracy of 63.7%, outperforming RL with verifiable rewards, RLAIF, and other baselines by 8 to 12%.

2 Related Work

Recent efforts to improve the quality and reliability of LLM explanations have primarily relied on RL paradigms based on subjective preference, most notably through Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017), Direct Preference Optimization (DPO) (Rafailov et al., 2023), or their structural variants. However, an explanation's perceived quality and its effect on decision quality are distinct evaluation targets. Buçinca et al. (2020) found that subjective measures of trust and preference for different explanation styles did not predict performance on the same decision-making tasks. Buçinca et al. (2021) showed that the most beneficial decision-support design received the least favorable subjective ratings. Moreover, optimizing for user preference can degrade human oversight, because users tend to favor fluent and agreeable outputs (Mitchell et al., 2026). These findings suggest that optimizing preferences can fail to lead to effective decisions.

Prior work also developed explanations targeted at impacting human decisions. Fukuchi & Yamada (2024) select combinations of explanations to encourage agreement with AI recommendations, using a predictive user model to estimate which combination will minimize the discrepancy between the user's anticipated decision and the AI's output. Similarly, Li & Yin (2024) use learned behavioral models to modify explanations toward targeted human decisions, strictly relying on the manipulation of static, score-based feature attributions. However, both frameworks lack an optimization objective that genuinely improves the user's task comprehension. Because they manipulate explanations merely to select a targeted behavioral response, rather than clarifying the underlying ground truth, they yield mixed empirical results and fail to improve independent human decision-making. ComplLLM (Guo et al., 2026) post-trains an LLM within a decision theoretic framework to extract signals that complement an existing recommendation, providing a direct precedent for optimizing language-model outputs according to decision value. Building on this perspective, we propose a decision theoretic RL-framework for explanation generation.

3 Framework: Calibrating Explanations on Human Decisions

Problem Setup A decision-maker observes an instance representation $x \in \mathcal{X}$, and must select an action $a \in \mathcal{A}$ under uncertainty about the ground truth state $y \in \mathcal{Y}$. The decision-maker also has access to an explanation $e \in \mathcal{E}$, generated by an LLM policy:

$$e \sim \pi_\theta(\cdot \mid x)$$

The quality of a decision is captured by a utility function $U : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ that assigns a real-valued score to a combination of action and state. We study a binary classification setting where $\mathcal{A} = \mathcal{Y} = \{0, 1\}$ and $U(a, y) = 1[a = y]$.

We are interested in optimizing explanations against decisions made by human decision-makers sampled from a target population. We use

$$a \sim P_H(\cdot \mid x, e)$$

to denote the action selected by a human decision-maker, where $P_H(\cdot \mid x, e)$ is the distribution of actions over the population given instance x and explanation e.

We would like to learn an explanation policy that maximizes the expected utility of the human decisions induced by the explanations. The target objective is

$$J_H(\theta) = \mathbb{E}_{(x,y)\sim D}\mathbb{E}_{e\sim\pi_\theta(\cdot|x)}\mathbb{E}_{a\sim P_H(\cdot|x,e)}\left[U(a,y)\right], \tag{1}$$

which is equivalent to

$$J_H(\theta) = \mathbb{E}_{(x,y)\sim D}\mathbb{E}_{e\sim\pi_\theta(\cdot|x)}\left[\sum_{a\in\mathcal{A}} P_H(a \mid x,e)\, U(a,y)\right]. \tag{2}$$

Modeling Human Decisions To directly optimize $J_H(\theta)$ requires oracle access to $P_H(\cdot \mid x, e)$, the distribution of human actions given instance x and explanation e. We therefore learn a human response model $S_\phi(a \mid x, e)$ on a dataset of observed human decisions to approximate $P_H(a \mid x, e)$.

Given S_ϕ, the predicted utility of an explanation on instance (x, y) is

$$\widehat{V}_\phi(x,e,y) = \sum_{a\in\mathcal{A}} S_\phi(a \mid x,e)U(a,y). \tag{3}$$

In our binary classification setting with $U(a, y) = \mathbf{1}[a = y]$, this reduces to

$$\widehat{V}_\phi(x,e,y) = S_\phi(a = y \mid x,e), \tag{4}$$

in other words, the predicted probability that a human who sees explanation e with instance x selects the correct action.

Policy Optimization Our target quantity is the simulator-predicted probability of a correct human decision,

$$q_\phi(x,e,y) = S_\phi(a = y \mid x,e).$$

Rather than use q_ϕ directly as the RL reward, we optimize a log-transformed surrogate,

$$r_\phi(x,e,y) = \max\left\{1 + \log q_\phi(x,e,y), -5\right\}. \tag{5}$$

Before clipping, the logarithm is strictly increasing in q_ϕ, so the surrogate preserves the ordering of candidate explanations by predicted human decision utility. The transformation places greater penalty on explanations for which the response model assigns very low probability to the correct human action. The lower bound limits the influence of extreme predictions from the learned response model. We optimize the explanation policy using this surrogate reward, with regularization toward a reference policy:

$$J_{\mathrm{RL}}(\theta) = \mathbb{E}_{(x,y)\sim D}\mathbb{E}_{e\sim\pi_\theta(\cdot|x)}\left[r_\phi(x,e,y) - \beta D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x)\|\pi_{\mathrm{ref}}(\cdot \mid x)\right) - \gamma\,\mathrm{len}(e)\right]. \tag{6}$$

4 Instantiating the framework on AI text detection

As shown in Figure 1, we implement the HD-RL framework for the human decision task of AI text detection, where people are faced with texts and must judge whether they were written by a human or an AI model (Russell et al., 2025; He et al., 2024). This process has five steps: Texts Collection, Explanation Generation, Human Decision Collection, Decision Simulator Training, and Reinforcement Learning. All prompts used during the experiments are provided in Appendix B.

4.1 Text Dataset Collection

We construct a dataset of mixed human and AI authored texts. Each instance in the dataset comprises a short text and a binary ground-truth label, where 1 indicates AI-generated content and 0 indicates human-written content. To collect the human-written source data, we initially scraped 10,000 texts

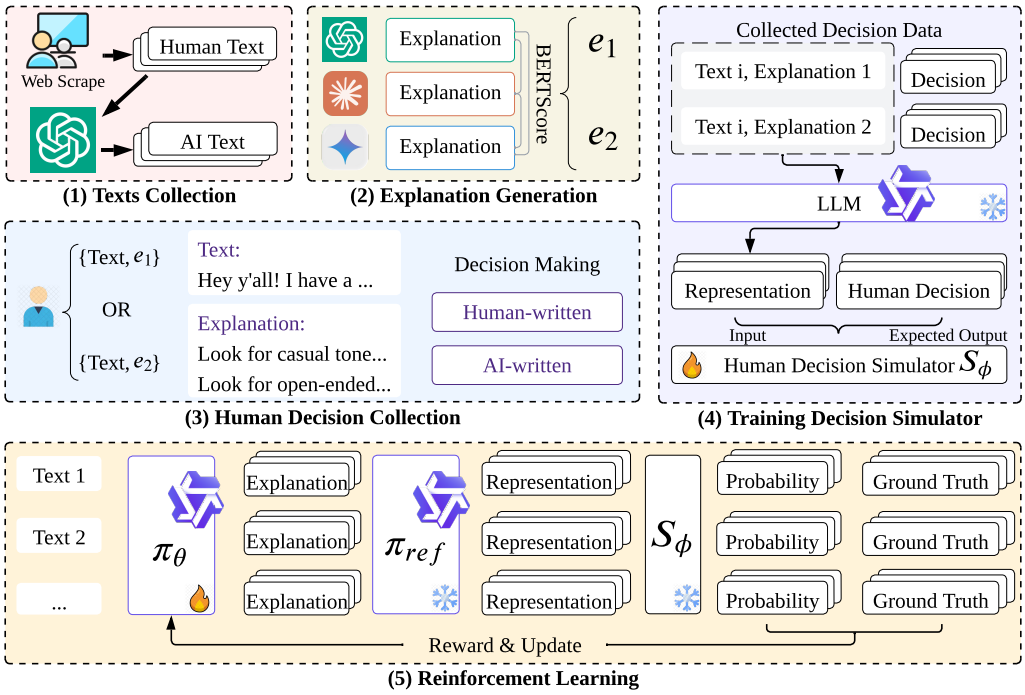


Figure 1: The overall architecture of the HD-RL framework. Modules marked with a fire icon represent trainable parameters, whereas those marked with a snow icon denote frozen models.

from Reddit published prior to November 30, 2022 (the release date of ChatGPT). To maintain participant engagement during the human decision collection stage, we applied a length filter, retaining only English texts containing between 50 and 250 words. This filtering process yielded a finalized set of 2,260 human-written texts. To construct the parallel AI-generated texts, we employed a two-stage LLM pipeline using GPT 5.4. For each human-written text, the model first generated a concise summary. We then prompted the LLM to rewrite the text based on the summary, requiring it to match the original word count. This process yielded a balanced final dataset comprising 4,520 total texts, which we partitioned into three subsets using a 60/25/15 split at the source-pair level. The original human text and its AI rewrite were always assigned to the same split. A summary of the dataset is given in the Appendix C.1.

4.2 Explanation Generation

LLM explanations often contain a prediction, which can lead users to accept the system's conclusion rather than evaluating the text themselves. To prevent this, the primary goal of our candidate explanations is to guide the user's attention to specific features without making the final decision for them. Inspired by the visual scaffolding techniques used to help humans spot AI-generated images (Kamali et al., 2024), we format our explanations as a structured set of localized feature indicators:

- Look for [specific linguistic or semantic feature]...
- Look for [structural, rhetorical, or thematic pattern]...

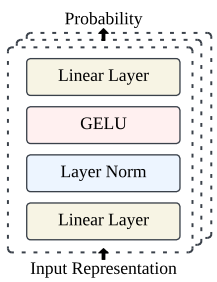
A detailed justification of this design choice and its advantages is provided in Appendix D.

For each text $x \in \mathcal{X}$, we generate a pool of candidate explanations using an ensemble of frontier LLMs $\mathcal{M} = \{M_1, M_2, M_3\}$ (GPT-5.4, Claude Opus 4.5, and Gemini 3.1):

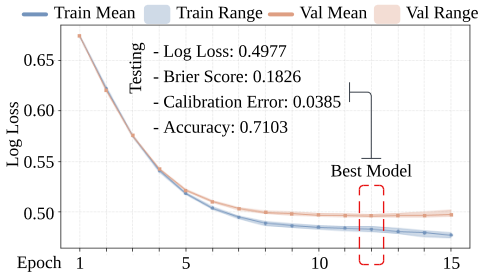
$$e^{(k)} \sim M_k(\cdot \mid x, \mathcal{P}_{\text{few-shot}}), \quad \text{for } k \in \{1, 2, 3\} \tag{7}$$

Metric	Value
Total Unique Participants	2,452
Total Decisions Collected	21,263
Unique Texts Evaluated	2,712
Average Trials per Participant	8.67
Average Decision per Pair	3.92
Initial Accuracy (Human Text)	**61.17%**
Initial Accuracy (AI Text)	**41.16%**
Overall Initial Accuracy	**51.28%**

(a) Human Decision Data Statistics



(b) Simulator Architecture



(c) Simulator Training Plot

Figure 2: Human Decision Data and Decision Simulator Training Report.

where $\mathcal{P}_{\text{few-shot}}$ is the prompt that contains expert-annotated demonstration exemplars. For each text x, this process yields a candidate set of explanations $\mathcal{C}(x) = \{e^{(1)}, e^{(2)}, e^{(3)}\}$. Because independent models sometimes generate highly overlapping explanations (e.g., describing very similar cues to look for), we eliminate redundant candidates. For each candidate pair $(e^{(u)}, e^{(v)}) \in \mathcal{C}(x) \times \mathcal{C}(x)$, we measure semantic similarity using token-level embedding alignment via BERTScore: $S(e^{(u)}, e^{(v)}) = \text{BERTScore}_{\text{F1}}(e^{(u)}, e^{(v)})$. We then keep the most semantically distinct subset $\mathcal{E}^*(x) = \{e_1, e_2\} \subset \mathcal{C}(x)$. Consequently, every text x is mapped to two distinct explanations (x, e_1) and (x, e_2), ensuring diverse diagnostic signals for the subsequent human decision collection stage.

4.3 Human Decision Collection

We collected human decisions on a public research website that hosts several media-forensics experiments and draws a general audience. Participants were unpaid volunteers who opted in from the site's landing page and gave informed consent before their first trial. Participants were identified only by an anonymous browser token. We collected no demographic information.

Each trial t presents a human decision-maker with a text $x_t \in \mathcal{X}$ alongside a designated explanation $e_t \in \mathcal{E}^*(x_t)$, and selects an action $a_t \in \{0, 1\}$, where 1 denotes AI text and 0 denotes human text. Upon submission of the participant's decision, the system reveals the ground-truth state $y_t \in \mathcal{Y}$ as feedback before moving to the next trial. We used session-based progress tracking to strictly guarantee that no participant was ever exposed to the same text twice, even across multiple visits. Furthermore, the interface prioritized serving instances (x, e) with the current lowest total number of collected human decisions. More detailed task information is in the Appendix E.

Dataset Summary Human data collection was conducted over a four-month period. As shown in Figure 2 (a), this stage collected 21,263 binary decisions from 2,452 unique participants for 2,712 unique texts (1,356 AI-generated and 1,356 human-written). Participants completed as many trials as they wished and could stop at any point. On average, each participant completed 8.7 decision trials; 5% completed more than 25. Each text-explanation pair received 3.9 decisions. Overall human decision accuracy using the unoptimized explanations was near-random at 51.3%, where participants achieved a 61.2% accuracy when identifying human-written texts, and 41.2% when identifying AI-generated texts.

4.4 Training Decision Simulator

For any given text-explanation pair (x_i, e_i), we obtain a set of human actions $\mathcal{H}_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,K_i}\}$, where $a_{i,j} \in \mathcal{A}$ and K_i represents the number of human decision-makers evaluating the i-th pair. We train an ensemble of five shallow MLP probes as a human decision simulator that takes an (x, e) pair and predicts the human decision distribution $P_H(a \mid x, e)$. Each MLP is trained with distinct random weight initializations and data permutations.

Conceptually, we need to learn a mapping function that generalizes from the observed (x, e) pairs in our dataset to arbitrary new (x, e) pairs generated during RL. Because x and e are natural language, this function requires a robust representation of the pair from which human responses can

be accurately predicted. A standard supervised approach might train a text encoder from scratch or fine-tune a language model to map the text directly to human actions. However, training a language model on 21K decision data risks overfitting. We use Qwen 3.6 as a frozen feature extractor. We extract the representation $\mathbf{h}_i \in \mathbb{R}^d$ of the (x, e) pair from the LLM's last layer residual stream, where $d = 5120$ denotes the hidden dimensionality. We then train a lightweight simulator to predict human decision probabilities based on the $\mathbf{h}_i$. Because this simulator only updates a small set of parameters, it converges quickly, avoids overfitting, and effectively generalizes to arbitrary new explanations (Jahanparast et al., 2026). As shown in Figure 2 (b), each individual MLP projects $\mathbf{h}_i$ into a lower-dimensional hidden space (256 dimensions), applies Layer Normalization to stabilize the input vectors, and uses a Gaussian Error Linear Unit (GELU) activation followed by Dropout ($p = 0.2$). A final linear projection and sigmoid activation yield the pre-ensemble probability. The simulator's final predicted probability, $\hat{p}_i = S_\phi(a = 1 \mid x_i, e_i)$, is the unweighted average of the outputs from the 5 MLPs. The ensemble parameters ϕ are optimized directly over the dataset using a Binary Cross-Entropy (BCE) loss function:

$$\mathcal{L}(\phi) = -\frac{1}{\sum_{i=1}^{N} K_i} \sum_{i=1}^{N} \sum_{j=1}^{K_i} \left[a_{i,j} \log(\hat{p}_i) + (1 - a_{i,j}) \log(1 - \hat{p}_i) \right], \tag{8}$$

where N is the total number of unique text-explanation pairs. This yields a calibrated continuous reward signal for the subsequent RL stage.

Decision Simulator Performance Before evaluating the decision simulator, we first establish its maximum achievable performance on our dataset. Because the model has no information about respondent identity, for any fixed (x, e) pair, the best it can do is predict the population action distribution. Therefore, we compare our simulator's performance to the best metrics that could be achieved by a model knowing only the (x, e) pair. Using the response probability for every observed pair gives an in-sample minimum achievable Log Loss (0.49) and Brier Score (0.17). Similarly, predicting the majority response for each pair establishes the maximum individual-decision accuracy (72.9%) for a model with access only to (x, e). To assess out-of-sample generalization, we evaluated the simulator on a 20% held-out test set, partitioned at the text level. Benchmarked against these limits, our simulator performs very well, as shown in Figure 2(c), achieving a predictive accuracy of 71.0%, a Log Loss of 0.50, and a Brier Score of 0.18, alongside an Expected Calibration Error (ECE) of 0.04. To confirm this predictive power stems from the explanations rather than the source text, we compare the simulator against a text-only ablation in Appendix F.1.

4.5 Reinforcement Learning

Our RL architecture comprises a trainable explanation policy π_θ, a frozen reference model π_{ref}, and the frozen decision simulator S_ϕ. Both π_θ and π_{ref} are initialized from the Qwen 3.6 27B model. The reference model π_{ref} serves a dual purpose: it enforces KL-divergence penalties to prevent policy collapse and acts as the representation extractor for the reward pipeline. In this stage, we use 25% of the pre-split dataset, ensuring the policy optimizes its explanation generation on a completely unseen distribution of texts. We optimize the policy π_θ using Group Relative Policy Optimization (GRPO). For a given input text x, the policy samples a group of $G = 8$ candidate explanations $\mathcal{E} = \{e_1, e_2, \ldots, e_G\}$. The advantage A_k for each candidate e_k is computed by standardizing its reward r_k against the mean and standard deviation of the group's rewards:

$$A_k = \frac{r_k - \mu(\mathbf{r})}{\sigma(\mathbf{r})}, \tag{9}$$

where $\mathbf{r} = \{r_1, r_2, \ldots, r_G\}$. To compute the reward r_k for a candidate explanation e_k, we execute a direct forward pass through our frozen simulator pipeline. We concatenate the text x and the generated explanation e_k, passing this prompt through π_{ref} to extract the 5120-dimensional representation $\mathbf{h}_k$. $\mathbf{h}_k$ is then routed to the frozen simulator S_ϕ, which outputs the predicted probability distribution of the human's actions. Consistent with the log-transformed surrogate defined in Equation 6, the reward is computed as:

$$r_k = \max\left(1 + \log q_\phi(x, e_k, y), -5.0\right), \tag{10}$$

where $\log q_\phi(x, e_k, y) = -\text{BCE}(\hat{p}_k, y) = y \log(\hat{p}_k) + (1-y)\log(1-\hat{p}_k)$. This mathematically maps the simulator's predicted probability of a correct human decision directly to the negative Binary Cross-Entropy, penalizing explanations that induce confident-but-wrong human actions.

5 EXPERIMENT

We conduct a human decision experiment on AI text detection, comparing decision quality with HD-RL to that of three baseline post-training methods.

5.1 BASELINES

Direct Preference Optimization on AI Feedback (DPO-RLAIF) The first baseline evaluates standard preference-based performance based on AI feedback. Following the explanation generation stage, each text x is paired with two explanations, e_1 and e_2. While gathering large human preference labels is expensive, recent research on Reinforcement Learning from AI Feedback (RLAIF) demonstrates that AI-generated preference labels can achieve alignment performance highly comparable with Reinforcement Learning from Human Feedback (RLHF) (Lee et al., 2023; Zheng et al., 2023). Leveraging this insight, we use Qwen 3.6 as an LLM judge to evaluate the pairs and select the preferred explanation, yielding a dataset of $(x, e_{chosen}, e_{rejected})$ tuples. We then optimize the policy model on these preferences using DPO.

Verifiable-Reward RL (VR-GRPO) The second baseline evaluates outcome-based RL performance. During training, we prompt the policy model to generate a joint output: an explanation $e \in \mathcal{E}$ followed by an autonomous action $a_{\text{LLM}} \in \mathcal{A}$. The reward is computed by comparing the LLM's action a_{LLM} directly against the ground-truth state $y \in \mathcal{Y}$. We assign a reward of $+1.0$ if $a_{\text{LLM}} = y$ and a penalty of -1.0 if $a_{\text{LLM}} \neq y$. The policy is optimized using GRPO.

LLM Decision Calibration The third baseline serves as a direct ablation of the Human Decision Collection stage, designed to test whether LLM simulations can be substituted for real human decision-making data. It replicates our framework exactly but replaces human decisions with synthetic LLM decisions. We prompt Qwen 3.6 to act as a human decision-maker, generating $21,263$ synthetic decisions over text-explanation pairs. The decision simulator S_ϕ is then calibrated entirely on this synthetic distribution.

Implementation Details All models, including the proposed framework and the respective baselines, are initialized using the Qwen 3.6 27B model (Qwen Team, 2026). To enable parameter-efficient training, we apply Low-Rank Adaptation (LoRA) to the policy models, configured with rank $r = 16$ and scaling factor $\alpha = 32$. Training is conducted across 4 NVIDIA A100 (80GB) GPUs. We optimize the models for 3 epochs using a learning rate of 1×10^{-5}. The batch size is set to 8 per device, with a gradient accumulation step of 4, yielding an effective global batch size of 128. During the GRPO exploration phase, the generation temperature is set to 1.0, and the policy model samples a group size of 8 candidate generations per input state. We apply a KL divergence penalty with a coefficient of $\beta = 0.04$. A length penalty of $\gamma = 0.001$ per token to encourage concise explanation. All prompts used in baselines are detailed in Appendix B.

5.2 RANDOMIZED CONTROLLED HUMAN EXPERIMENT

After training, we conducted a randomized between-subjects repeated measures human decision experiment to evaluate the downstream utility of the generated explanations. We recruited 693 human participants from the same target population. Upon beginning the experiment, participants were randomly assigned to one of four experimental conditions (HD-RL, LLM_decision, DPO-RLAIF, or VR-GRPO). Each participant then evaluated a sequence of texts drawn from the 15% pre-split dataset (678 unique texts: 339 AI-generated and 339 human-written) paired with the explanations generated by their assigned condition's policy. In total, we collected 7,455 human trials. While

random assignment led to unequal total participant counts across the conditions, the average engagement remained highly consistent: participants completed an average of 10.8 trials in HD-RL, 11.6 in LLM_decision, 10.2 in DPO-RLAIF, and 10.8 in VR-GRPO. For a detailed breakdown of the human decision dataset, refer to Table 1 in the Appendix.

Because raw empirical metrics (such as accuracy or F1-scores) can be confounded by the inherent difficulty of specific texts or baseline differences in participant skill, we rely on model-derived estimates to report performance. We fit a joint Generalized Linear Mixed Model (GLMM) with a logistic link function across the entire deployment dataset. The mixed-effects model is defined as:

$$\text{logit}(P(a = y)) = \beta_0 + \beta_{\text{cond}} + \beta_{\text{AI}} \cdot y + \beta_{\text{cond}\times\text{AI}} + \beta_{\text{trial}} \cdot \text{trial_num_c} + u_{\text{participant}} + u_{\text{text}}, \quad (11)$$

where $P(a = y)$ represents the probability that the human decision-maker selects the correct action. Fixed effects include β_0, which serves as the global intercept, representing the reference condition (HD-RL) on human-written texts, while β_{cond} captures the log-odds offset of each baseline method relative to HD-RL. The term $y \in \{0, 1\}$ denotes the ground truth (1 for AI, 0 for Human), making β_{AI} the log-odds offset of a correct decision when evaluating AI-generated texts versus human-written texts for the reference condition. $\beta_{\text{cond}\times\text{AI}}$ is an interaction term that allows the AI-vs-Human performance gap to differ by method; we include this term based on observations from a pilot study that explanations from some methods, such as DPO-RLAIF, seemed to favor suggesting human ground truth, suggesting explanation effects could vary based on ground truth . Additionally, trial_num_c is a centered continuous variable controlling for learning effects. Finally, the model incorporates two random effects: $u_{\text{participant}} \sim \mathcal{N}(0, \sigma_p^2)$ acts as a nested random intercept for each participant (as users only participated in one condition), and $u_{\text{text}} \sim \mathcal{N}(0, \sigma_t^2)$ serves as a crossed random intercept, because the 678 texts are shared across all conditions.

6 Results

Figure 3 illustrates the model-derived estimates of human performance across all experimental conditions in the final deployment dataset. Comprehensive statistical results, including the full GLMM fixed-effect coefficients, estimated marginal accuracies, and odds-ratio contrasts by ground truth, are provided in Appendix G.

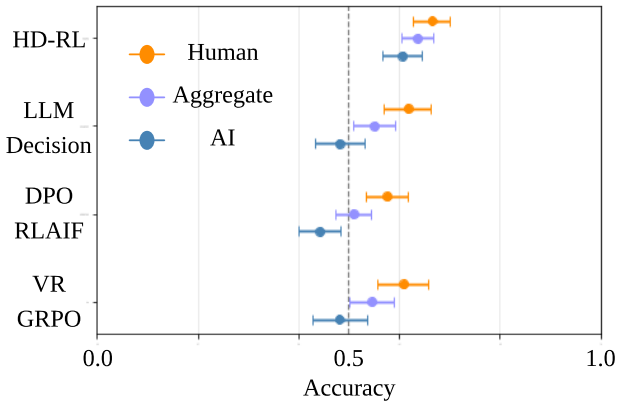


Figure 3: GLMM-Estimated Human Accuracy across Experimental Conditions. This plot displays the model-derived human accuracy estimates for each alignment framework, complete with 95% Confidence Intervals. For each condition, the aggregate performance (purple) is staggered alongside the disaggregated performance on human-written texts (orange) and AI-generated texts (blue).

6.1 Improvement in Human Decision Making

We report model-estimated average accuracies by condition and conditional on ground truth source (AI or human), with 95% confidence intervals in Figure 3. HD-RL yielded an estimated accuracy of 63.7% (95% CI: [60.4%, 66.8%]), corresponding to a raw empirical accuracy of 65.6%. This outperforms the best-performing baseline, LLM_decision, which only reached an estimated accuracy of 55.1% (95% CI: [50.9%, 59.2%]). We observed similar improvements across both text classes, in contrast to asymmetric vulnerabilities seen in baselines. While all baseline models yielded low accuracy on AI-generated texts (driving raw accuracy down

to between 45% and 49%), HD-RL explanations elevate human accuracy on AI texts to 63.0%. These results suggest that the alternative reward signals used by other methods do not lead to explanations that adequately discriminate AI from human texts for human decision-makers.

6.2 Divergence Between Human and Synthetic Decision Distributions

The LLM decision baseline serves as an ablation of the human decision collection phase, training the decision simulator entirely on synthetic LLM decisions. While this approach performed better than other baselines (yielding a raw overall human accuracy of 55.3% and a GLMM-estimated accuracy of 55.1%), it still fell over 8% short of HD-RL framework.

To understand this performance gap, we compare the underlying decision-making patterns present in the human and synthetic decision datasets. As detailed in Appendix C.6, when an LLM acts as a proxy participant, it achieves an initial detection accuracy of 76.4%. Human participants, on the other hand, achieve near-random baseline accuracy. We hypothesize that this discrepancy exists because the LLM already possesses the latent knowledge required to identify AI-generated text. While human decisions are inherently noisy and occasionally conflicting, they encode critical behavioral cues regarding which specific explanatory features help people to make better decisions and which do not. For instance, as demonstrated in Appendix C.5, human participants achieve higher accuracy when explanations focus on specific, class-appropriate lexical phrases.

6.3 Asymmetries in Explanation Generation and Selection

As detailed in Appendix C.3, we observe an asymmetry in the generated explanations, which frequently highlight features associated with human authorship. Although our explanation format strictly avoids revealing explicit predictions, the wording can still imply a human versus an AI origin by using terms more strongly associated with perceptions of humans versus machines. For example, the following example is an AI-generated text:

> *"Starting a new embedded systems job soon and honestly I'm pumped but pretty nervous too. Feel a bit underqualified and could really use some advice from anyone who's been there..."*

An explanation such as, ``Look for neatly balanced emotional contrasts like 'pumped but pretty nervous too,''' highlights structural symmetry, implicitly suggesting AI authorship. Conversely, an explanation stating, ``Look for informal emotive language like 'pumped' and 'nervous,''' focuses on conversational imperfection, implicitly suggesting human authorship. During the initial explanation generation stage, we observed that LLMs frequently generated explanations focusing on human informality, regardless of the text's actual label.

Furthermore, in the DPO-RLAIF baseline, where an LLM acts as a judge to select preference pairs, this asymmetry is amplified. As detailed in Appendix C.4, even when evaluating AI texts, the LLM judge selected human-implying explanations in over 83% of cases. This lexical-focus analysis is one possible reason for the class asymmetry observed in downstream performance. Human participants evaluating texts with DPO-RLAIF explanations achieved an overall accuracy of just 51.9% (GLMM-estimated accuracy: 51.0%, 95% CI: [47.5%, 54.5%]), and their detection accuracy on AI texts was 45.7%.

7 Conclusion

In this paper, we introduced HD-RL, a decision-theoretic RL framework that aligns LLM explanation with downstream human utility. By applying HD-RL to the Qwen 3.6 model, we generate explanations that help people make good decisions, achieving a GLMM-estimated human accuracy of 63.7% in a randomized between-subjects experiment with repeated measures and outperforming RLVR, RLAIF, and other baselines by over 8%. Ultimately, this work highlights the need for grounding LLM outputs in human action to ensure their usefulness in human-AI collaborative settings. We release our human decision dataset and trained model for future work.

REFERENCES

James O Berger. Statistical decision theory. In *The New Palgrave Dictionary of Economics*, pp. 1–6. Springer, 1987.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 454–464, 2020. doi: 10.1145/3377325.3377498. URL https://arxiv.org/abs/2001.08298.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021. doi: 10.1145/3449287. URL https://arxiv.org/abs/2102.09692.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122 (25):e2412015122, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Tom Duenas and Diana Ruiz. The risks of human overreliance on large language models for critical thinking. *Research Gate*, 2(26002.06082), 2024.

Yosuke Fukuchi and Seiji Yamada. User decision guidance with selective explanation presentation from explainable-AI. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2024. URL https://arxiv.org/abs/2402.18016.

Ziyang Guo, Yifan Wu, Jason Hartline, Kenneth Holstein, and Jessica Hullman. ComplLLM: Fine-tuning LLMs to discover complementary signals for decision-making, 2026. URL https://arxiv.org/abs/2602.19458.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 2251–2265, 2024.

Erfan Jahanparast, Zhiqing Hong, and Serina Chang. What do large language models know about opinions? In *International Conference on Learning Representations*, volume 2026, pp. 146105–146128, 2026.

Negar Kamali, Karyn Nakamura, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh. How to distinguish ai-generated images from authentic photographs. *arXiv preprint arXiv:2406.08651*, 2024.

Eunhye Kim, Kiroong Choe, Minju Yoo, Sadat Shams Chowdhury, and Jinwook Seo. Beyond tools: Understanding how heavy users integrate llms into everyday tasks and decision-making. *arXiv preprint arXiv:2502.15395*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Peiming Li, Zhiyuan Hu, Shiyu Li, Xi Chen, and Yang Tang. Aligning deep implicit preferences by learning to reason defensively. In *International Conference on Learning Representations*, volume 2026, pp. 19024–19045, 2026.

Zhuoyan Li and Ming Yin. Utilizing human behavior modeling to manipulate explanations in AI-assisted decision making: The good, the bad, and the scary. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/08bd07d567d77d6dd8d82a4474706a5e-Abstract-Conference.html.

Margaret Mitchell, Avijit Ghosh, and Samir Passi. Ai agents push humans out of the loop. *arXiv preprint arXiv:2608.23642*, 2026.

Cláudia Vale Oliveira, Nelson Zagalo, Filipe Silva, Anabela Brandão, Syeda Faryal Hussain Khurrum, and Joaquim Santos. Plausibility as failure: How llms and humans co-construct epistemic error. *arXiv preprint arXiv:2512.16750*, 2025.

Shramay Palta, Peter A Rankel, Sarah Wiegreffe, and Rachel Rudinger. Arguments that alter minds: Llm rationales sway human (and llm) notions of plausibility. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13132–13152, 2026.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5342–5373, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Linghao Wang, Zheyuan Jiang, Chenke Hu, Jun Zhao, Zheng Zhu, Xiqun Chen, Ziyi Wang, Tianming Liu, Guibing He, Yafeng Yin, et al. Comparing ai and human decision-making mechanisms in daily collaborative experiments. *Iscience*, 28(6), 2025.

Qile Wang, Prerana Khatiwada, Avinash Chouhan, Ashrey Mahesh, Joy Mwaria, Duy Duc Tran, Kenneth E Barner, and Matthew Louis Mauriello. " the explanation makes sense": An empirical study on llm performance in news classification and its influence on judgment in human-ai collaborative annotation. *arXiv preprint arXiv:2602.19690*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A LIMITATION

While the goal of HD-RL is to improve human decision accuracy, our implementation relies on a logarithmic surrogate reward. Specifically, rather than directly optimizing the continuous expected accuracy $q_\phi(x, e, y) = S_\phi(a = y \mid x, e)$ predicted by the simulator, we optimize the transformed surrogate $1 + \log(q_\phi(x, e, y))$. This logarithmic transformation preserves the rank ordering of candidate explanations for a fixed instance prior to clipping, it inherently penalizes low-q_ϕ explanations more heavily than a linear reward would. Future work should compare this surrogate against the direct optimization of predicted expected utility.

B PROMPT USED

Prompt Used in Data Collection

Prompt for Summarization:

Summarize the following social media post with 1-3 sentences. Summarize strictly on the core intent or question of the post. Start the response immediately with the summary content. Strictly omit all introductions, meta-commentary, and conversational fillers.

Post content:

{text}

Prompt for Rewrite:

You are a highly active, authentic Reddit user tasked with rewriting a summary into a casual social media text post.

CRITICAL REQUIREMENT: Your rewrite must be approximately {target_word_count} words long.

To ensure the text reads like a real human posting on a forum and not an AI generation, strictly follow these rules:

1. Dive straight into the topic like you are speaking to a friend or dropping into an ongoing community discussion.

2. Do NOT use bullet points, numbered lists, emojis, generic language, or an overuse of em-dashes. Do Not list words in groups of three.

Summary to rewrite:

{summary}

Prompt Used in Explanation Generation

Task: Analyze the provided Social Media post for linguistic markers of AI or human authorship.

Constraints:

1. Start the response IMMEDIATELY with the bulleted list.

2. Do NOT provide an introduction, a final verdict, or a summary.

3. Do NOT group the bullet points into sub-headings (keep it a flat list).

4. Every bullet point MUST follow one of these two exact templates:

- "Look for [linguistic category or type of phrasing]."

- "Look for [linguistic category or type of phrasing] like '[short example]'."

5. When providing examples, you may list multiple words separated by commas (e.g., like "word 1", "word 2").

6. Do NOT use bolding, italics, or any special Markdown formatting.

7. Limit the output to exactly 2 to 6 bullet points.

Examples:

Post:

‘‘‘

I've always felt a deep connection to Taurus—something quiet, steady, and truly special. I used to think the zodiac signs were all about flashy personalities or dramatic emotions, but Taurus? They're the ones who show up without fanfare. I've watched friends go through tough times, and every time, a Taurus friend has been there—calm, grounded, offering a warm cup of tea or just sitting in silence with you, listening. They don't need to fix everything. They just show up with presence. That's rare.

Taurus isn't loud or flashy. They don't chase attention or try to be the center of the room. Instead, they build real, lasting connections. They're loyal—not just in words, but in action. When you're struggling, they don't offer quick fixes. They offer stability. They offer comfort. And that kind of emotional support? It's not just helpful—it's healing.

I've seen how Taurus people nurture their loved ones with patience and care. They take time to understand what others need, not what they want. They're dependable, grounded, and deeply kind. In a world that often values speed and intensity, Taurus brings balance. They remind us that strength isn't always loud—it can be quiet, consistent, and deeply human.

I've realized something: Taurus isn't just underrated. They're essential. They're the quiet pillars in our lives—steady, patient, and full of genuine warmth. I don't just appreciate them. I value them. And if you've ever had a Taurus friend, you know—this sign isn't just a zodiac sign. It's a gift. I'm so proud to stand beside them.

”’

Explanation:

- Look for em dashes.

- Look for descriptions followed by elaboration, setting a familiar rhythm to the sentences.

- Look for reasonable length paragraphs that tend to be consistent in length.

Post:

“‘

Taurus season started yesterday, and so I wanted to make a post dedicated to appreciating Taurus since I feel like Taurus is never really talked about and is very underrated. I find that I come across a lot of Taureans (probably cause I'm a Taurus descendant) and almost all of them I meet I find them to be nothing short of amazing. I believe that Taurus is the kindest and loyalest sign out of the zodiac, and they're very caring as well. Taurus offers a stable support system that you can certainly rely on, they always offer a hand to those they care about. They're great at advice, and can help you see the beauty in life's most difficult situations. They will always be there to encourage you to keep pushing forward even when your life seems to be getting off track, and they do this in a very gentle, caring and nurturing way. If you're lucky enough to be friends with them you'll know that they offer a ride or die type friendship, and they will always be thoughtful of you. You will also be aware that the stereotype of them being boring is far from true, they will surely have you entertained even in the most mundane situations. They are as loyal as they come and will not only fend for you in your presence but in your absence as well. They offer genuineness, stability, loyalty and beauty to your life and I very much appreciate all the Taureans I've had the pleasure to befriend.

”’

Explanation:

- Look for run on sentences and use of because to string together clauses.

- Look for colloquial expressions like “nothing short of X”.

- Look for non-descript positive words like “great” and “amazing.”

- Look for overly long paragraphs or blocks of text.

Post:

“‘

{text}

”’

Explanation:

Prompt Used in Training Human Simulator

Prompt for Extracting Representation:

You are tasked with detecting whether a given text is AI-generated or human-written. Below is the text, followed by an explanation hint that may help in your prediction.

Text:

{text_content}

Explanation Hint:

{explanation_content}

Based on the text and the explanation hint, predict whether the text is AI-generated or human-written.

Output 1 for AI-generated or 0 for human-written. Output only the number 0 or 1.

Prompt Used in Reinforcement Learning

Task: Analyze the provided text for linguistic markers of AI or human authorship.
Constraints:
1. Start the response IMMEDIATELY with the bulleted list.
2. Do NOT provide an introduction, a final verdict, or a summary.
3. Do NOT group the bullet points into sub-headings (keep it a flat list).
4. Every bullet point MUST follow one of these two exact templates:
- "Look for [linguistic category or type of phrasing]."
- "Look for [linguistic category or type of phrasing] like '[short example]'."
5. When providing examples, you may list multiple words separated by commas (e.g., like "word 1", "word 2").
6. Do NOT use bolding, italics, or any special Markdown formatting.
7. Limit the output to exactly 2 to 6 bullet points.

Text:
‘‘‘
{text}
’’’

Can you provide an explanation that tells the reader what to focus on?

Explanation:

Prompt Used in LLM Decision Baseline

You are evaluating a text to determine if it was written by an AI or a Human.
Text:
{text_content}
Explanation hint:
{explanation_content}
Based on the text and the explanation, make a decision.
Respond with EXACTLY one word: either 'AI' or 'Human'.

Prompt Used in DPO-RLAIF Baseline

You are an expert evaluator of AI vs. Human text detection heuristics.
I will show you a text and 2 different explanations that describe linguistic markers for determining whether the text is AI-generated or human-written.
Your task: Select which explanation MORE effectively captures the key linguistic markers that would help detect whether this text is AI or human.
Consider these criteria:
1. Clarity: Is the explanation clear and specific?
2. Relevance: Does it identify markers actually present in the text?
3. Usefulness: Would this explanation be most helpful for classification?
4. Distinctiveness: Does it capture markers unique to AI or human writing?
5. Coverage: Does it capture more relevant linguistic features?
Text:
‘‘‘
{text}
’’’

Explanation 1:
{exp_1}
Explanation 2:
{exp_2}
Respond with ONLY the number (1 or 2) of the better explanation, followed by a brief 1-2 sentence justification. Format:
[NUMBER]

[Justification]

Prompt Used in VR-GRPO Baseline

Task:
Analyze the provided social media post for linguistic markers of AI or human authorship.
Constraints:
1. Start the response IMMEDIATELY with the bulleted list of linguistic markers.
2. Do NOT provide an introduction, a final verdict summary, or other preamble.
3. Do NOT group the bullet points into sub-headings (keep it a flat list).
4. Every bullet point MUST follow one of these two exact templates:
- "Look for [linguistic category or type of phrasing]."
- "Look for [linguistic category or type of phrasing] like '[short example]'."
5. When providing examples, you may list multiple words/phrases separated by commas.
6. Do NOT use bolding, italics, or any special Markdown formatting.
7. Limit the explanations to exactly 2 to 6 bullet points.
8. After the bullet points, end with exactly this line: "JUDGMENT: AI" or "JUDGMENT: Human" (no extra text after this).
Post:
‘‘‘
text
’’’
Explanation:

C Explanations and Baselines Analysis

We conducted a comprehensive analysis of the text and human decision collection stages.

C.1 Dataset Distribution

The texts collection yielded a balanced final dataset comprising 4,520 total texts, which we partitioned into three subsets using a 60/25/15 split. The largest partition (60%) is dedicated to the decision simulator training. These texts will go through the explanation generation and human decision collection stages. A second partition (25%) is reserved for the reinforcement learning stage, providing an isolated environment to update the policy model. Finally, the remaining 15% is for the deployment stage, serving as an unseen set to evaluate the optimized model.

As illustrated in Figure 5(a), the distribution of human text lengths and AI text lengths are highly aligned. Both distributions are strictly bounded between 50 and 250 words, and the overall frequency trends of two texts are similar too.

C.2 Initial Human Accuracy Asymmetry

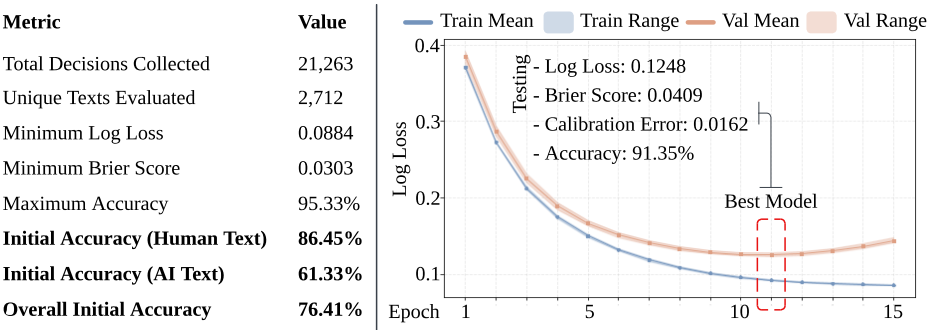
As discussed in main draft, the original explanations from closed-source LLMs achieved a near-random overall accuracy of 51.3%. We detailed the breakdown of human decision accuracy by explanation source in Figure 5(b). Across all three frontier models (GPT, Claude, and Gemini), the generated explanations were more helpful in identifying human-written texts than AI-generated texts. Specifically, explanations generated by GPT, Claude, and Gemini yielded human accuracy rates of 0.66, 0.59, and 0.56, respectively, when the ground truth was human. Conversely, when the underlying text was AI-generated, people only achieved accuracy of 0.37, 0.42, and 0.47, respectively.

C.3 Lexical Appropriateness Analysis

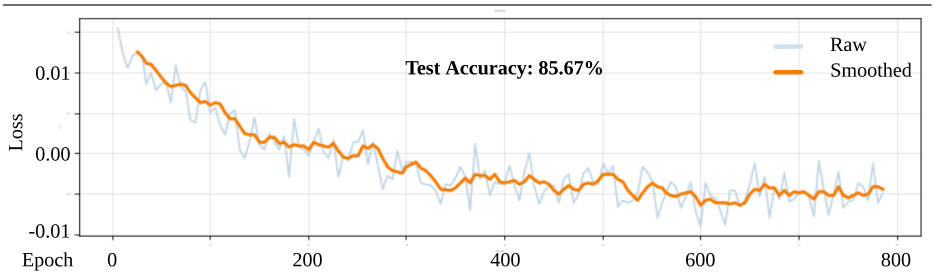
We conducted a lexical analysis to determine whether the generated explanations were appropriate for the given text label. In our project, even though the explanation doesn't explicitly reveal the

Metric	Value
Total Decisions Collected	21,263
Unique Texts Evaluated	2,712
Minimum Log Loss	0.0884
Minimum Brier Score	0.0303
Maximum Accuracy	95.33%
Initial Accuracy (Human Text)	**86.45%**
Initial Accuracy (AI Text)	**61.33%**
Overall Initial Accuracy	**76.41%**

(a) LLM Decision Data Statistics



(b) Simulator Training Plot



(c) VR-GRPO Training Loss Over Time

Figure 4: LLM Decision Baseline Simulator Calibration and VR-GRPO Training Report. (a) Data statistics for the LLM-generated synthetic decisions. (b) Training and testing performance of the simulator, demonstrating near-optimal convergence on the LLM's decision distribution. (c) VR-GRPO training loss over time, along with the test accuracy of 85.7%.

model's prediction, its wording can imply a certain label. We identified two dictionaries of label-specific indicator phrases:

- `HUMAN_INDICATORS`: Terms explicitly pointing out linguistic flaws or organic writing patterns (e.g., "imperfection", "grammatical", "awkward", "error", "mistake", "typo", "misspell").
- `AI_INDICATORS`: Terms highlighting algorithmic uniformity (e.g., "balanced", "carefully balanced", "calibrated", "smooth", "polished", "structured").

We then evaluated the explanations to see if they contained indicators appropriate to the ground-truth label of the text they described. As shown in Figure 5(c), 74.4% of the explanations generated for human texts appropriately focused on human-specific indicators. However, only 52.8% of the explanations generated for AI texts successfully utilized AI-specific indicators.

C.4 Asymmetry in LLM Judge Preferences

We analyzed the selection behavior of the LLM judge utilized in our DPO-RLAIF baseline. In this setup, the LLM is tasked with selecting the preferred explanation from a candidate pair. As illustrated in Figure 5(d), this process reveals a strong distributional asymmetry. When evaluating human-written texts, 91.1% of the explanations preferred by the LLM judge focus on human-associated lexical features. However, when evaluating AI-generated texts, 83.6% of the chosen explanations still focus on human-associated features. This observational pattern indicates that the preferences of the autonomous LLM judge diverge from the explanatory features that correlate with balanced human decision accuracy across both classes.

C.5 Empirical Grounding of the Simulator Reward

By analyzing human decision accuracy specifically on AI-generated texts in Figure 5(e), we observe a strong association between the explanation's lexical focus and downstream human performance.

When participants are provided with an AI-focused explanation for an AI text, average empirical accuracy is 65.3%. However, for AI texts paired with human-focused explanations, average accuracy is only 37.3%. While this observational contrast does not isolate the causal effect of the explanation from texts, it highlights empirical discrepancy associated with lexical focus. Because HD-RL framework optimizes the policy against the calibrated human decision simulator rather than an autonomous LLM judge, the policy model is rewarded for navigating these empirical lexical patterns.

C.6 Simulator Analysis for LLM Decision Baseline

In LLM Decision baseline, rather than optimizing the policy against human decisions, we replaced people with an LLM, prompting it to act as human participants. We prompted the LLM to generate exactly 21,263 synthetic binary decisions across 2,712 unique texts. As detailed in Figure 4(a), the LLM demonstrated a higher accuracy for AI text detection than actual humans. The LLM achieved an overall initial accuracy of 76.4%. This contrasts with the near-random 51.3% overall accuracy exhibited by real humans, proving that the LLM's synthetic decision distribution (P_{LLM}) diverges substantially from the human distribution (P_H). Similar to our human decision simulator, we establish the maximum achievable performance for a model trained on this synthetic dataset. Because a predictive model with access only to the (x, e) pair lacks information about the specific "respondent" (in this case, the LLM's sampling variance), the best it can do is predict the empirical action distribution for each instance. Using the empirical response probability for every observed (x, e) pair establishes an in-sample minimum achievable Log Loss of 0.09 and a Brier Score of 0.03. Furthermore, predicting the majority synthetic response for each instance caps the maximum empirical individual-decision accuracy at 95.3%.

To assess out-of-sample generalization, we evaluated the simulator on a 20% held-out test set, partitioned strictly at the text level. Benchmarked against these limits, the trained LLM decision simulator achieved highly optimized performance. As illustrated by the training plot in Figure 4(b), the final simulator yielded a Log Loss of 0.13, a Brier Score of 0.04, and an Expected Calibration Error (ECE) of 0.02.

D Design and Justification of the Explanation Format

In this project, our objective is to provide useful diagnostic explanations without inducing overreliance or fatigue. Our explanations format looks like:

- Look for [specific linguistic or semantic feature]...
- Look for [structural, rhetorical, or thematic pattern]...

This specific format was selected for four primary reasons. (1) Reading complex explanations introduces cognitive fatigue, often causing users to ignore the text entirely. Our format minimizes reading time, presenting the user with straightforward criteria. (2) A common failure mode in human-AI collaboration is when users focus more on the explanation rather than the source text. Our explanation forces the human to redirect their attention back to the text to verify if the suggested pattern actually exists. (3) Our explanation never reveals a decision to the user (e.g., it never says, "This text is AI-generated because..."). Stating a decision introduces severe anchoring bias, where humans unconsciously align their belief with the AI's stated belief. Therefore, we let the human control the final decision, preserving ultimate human agency. (4) Our format is highly controllable. LLMs can easily follow this syntactic structure. Furthermore, we make downstream automated processes, such as BERTScore filtering, much more accurate and computationally stable. For specific examples of texts and explanations, refer to Appendix H.

E Website Design and Quality Assurance

In this study, we designed a custom web platform to collect human decision data.

Upon visiting the platform, participants are first required to review and accept a Terms of Use agreement outlining the academic nature of the study and data privacy protocols. We also show user a project description:

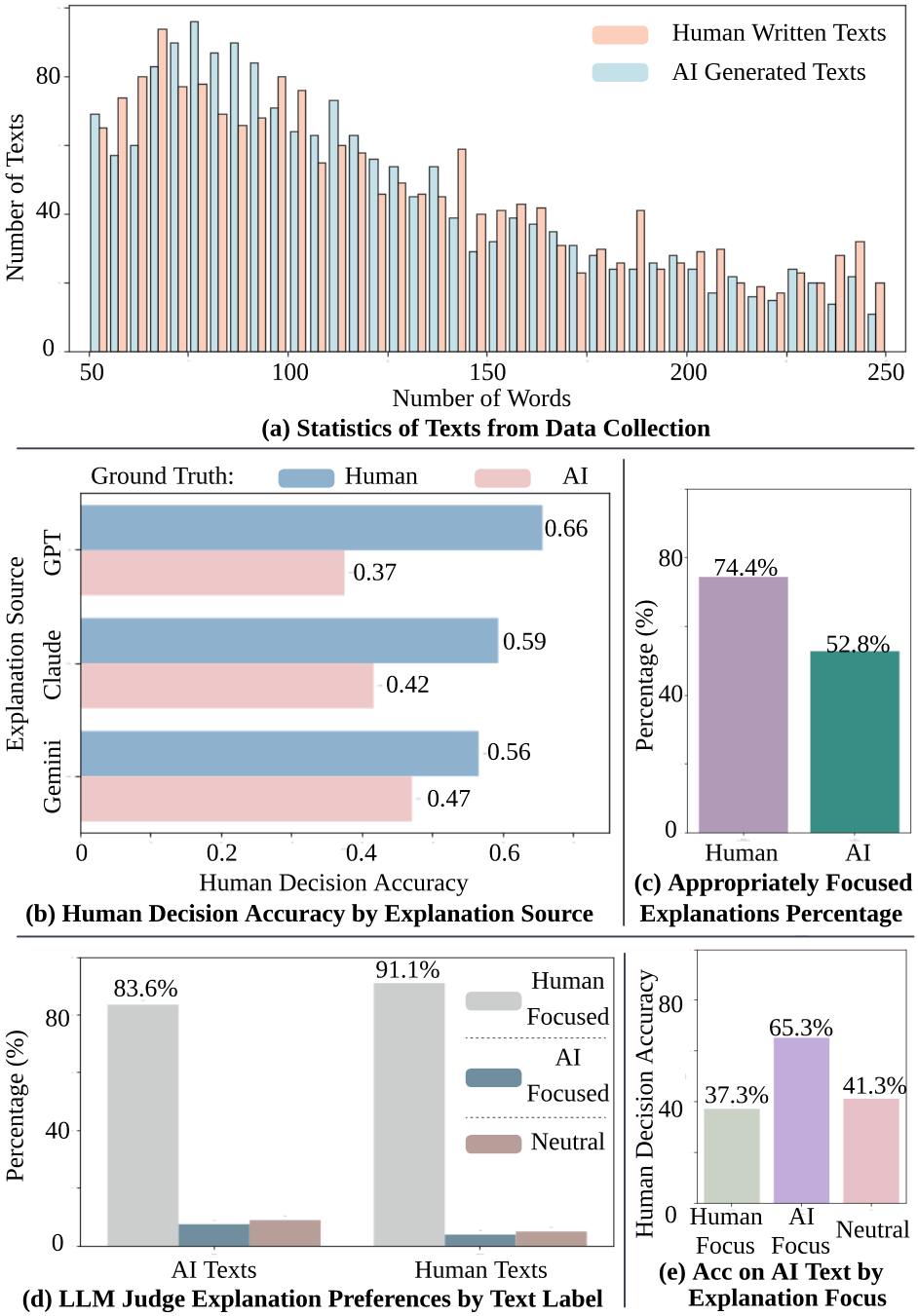


Figure 5: Analysis of the text dataset, human decisions, and explanations. (a) Text length distributions for human and AI texts demonstrate structural parity. (b) Human decision accuracy broken down by explanation source reveals a severe skew toward human text detection. (c) The percentage of explanations containing vocabulary appropriate for the ground-truth text label. (d) The LLM judge exhibits a massive bias toward human-focused explanations regardless of the ground-truth label. (e) Human decision accuracy on AI texts demonstrates that AI-focused explanations are required for high human accuracy.

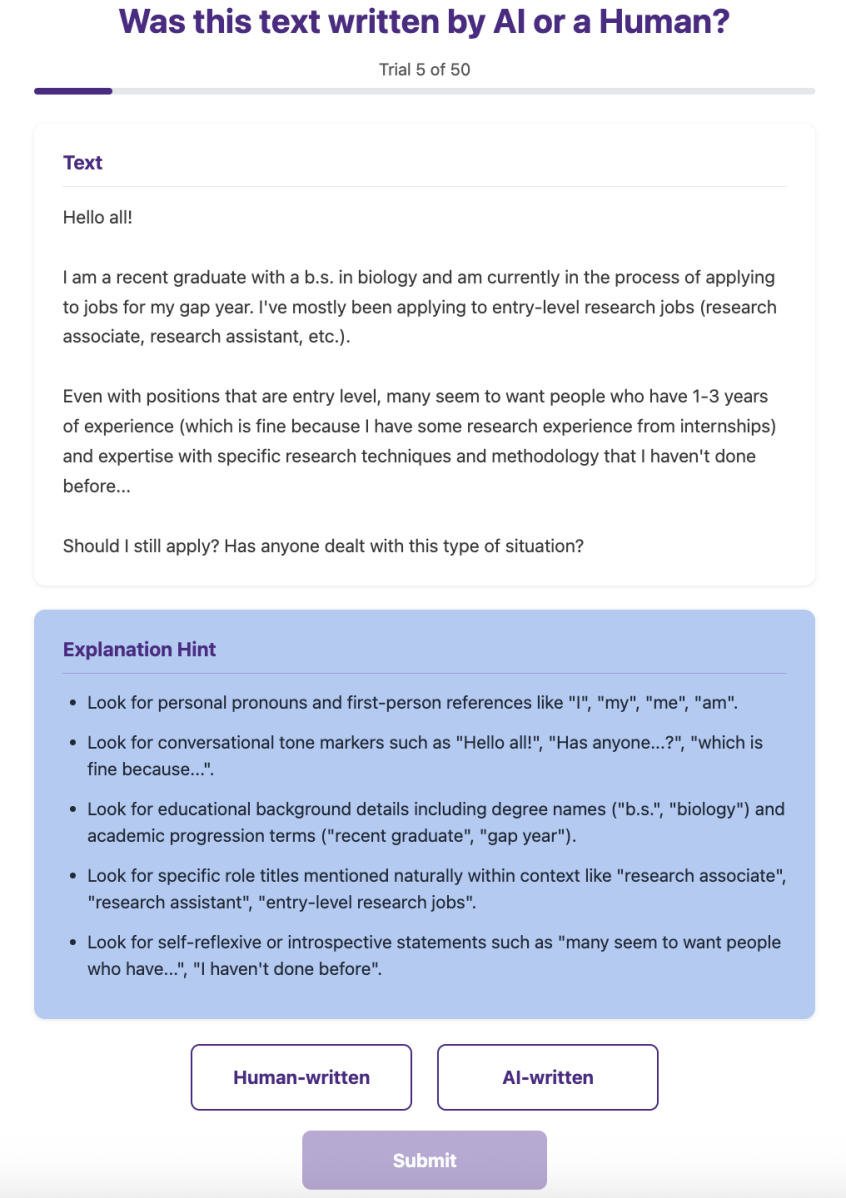


Figure 6: Website UI Interface.

```
In this experiment, you will be presented with a series of short
social-media texts.  Each text was written either by a person
or generated by AI. We ask you to read each text carefully and
share whether you believe it was written by a Human or by an AI.
An ``explanation hint'' will be shown alongside each text to help
you decide.
```

To minimize onboarding friction, the platform does not require account registration. Instead, we utilize browser cookies to assign and track unique participant IDs over time. As shown in Figure 6, the UI is deliberately designed to minimize the human cognitive load. The main workspace is vertically split into two parts. The upper container displays the text against a white background to prioritize readability. Immediately below, the explanation is displayed in a blue container, visually separating from the text. At the bottom of the interface, participants lock in their decision using the "Human-written" or "AI-written" buttons before clicking the "Submit" button.

F Decision Simulator Details

F.1 Decision Simulator Ablation

Because the HD-RL framework relies on GRPO to reward different explanations for a fixed text, the decision simulator S_ϕ must be sensitive to the specific explanation e, rather than simply predicting the human action distribution $P_H(a \mid x)$ based on the difficulty of the text x. To isolate the marginal predictive value of the explanations, we conducted an ablation study by training a baseline simulator exclusively on the source texts. Removing the explanation e degraded predictive accuracy from 71.0% to 60.8%, while Log Loss increased from 0.50 to 0.62, confirming that our simulator effectively captures the explanation-level fidelity.

F.2 Decision Simulator Evaluation Metrics

When evaluating the performance of the decision simulator S_ϕ, we treat the target action as $a'_i \in \{0, 1\}$. The simulator outputs a continuous predicted probability $\hat{p}_i = S_\phi(a = 1 \mid x_i, e_i)$. We evaluate the simulator's predictive fidelity using the following metrics:

Log Loss (Binary Cross-Entropy) Log Loss measures the distance between the predicted probability distribution and the discrete target actions, heavily penalizing confident but incorrect predictions. This serves as our primary surrogate reward signal:

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[a'_i \log(\hat{p}_i) + (1 - a'_i)\log(1 - \hat{p}_i)] \tag{12}$$

Brier Score The Brier Score calculates the mean squared difference between the predicted probability and the actual outcome, providing a bounded measure (0 to 1) of probability forecast accuracy:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - a'_i)^2 \tag{13}$$

Expected Calibration Error (ECE) To compute Calibration Error, we partition the N predictions into M equally spaced probability bins $\{B_1, \ldots, B_M\}$. ECE measures the expected absolute difference between a bin's empirical action 1 frequency and its mean predicted probability. A perfectly calibrated simulator yields an ECE of 0:

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}|\text{freq}(B_m) - \text{conf}(B_m)| \tag{14}$$

where $\text{freq}(B_m)$ is the empirical proportion of observed action 1 decisions in bin m, and $\text{conf}(B_m)$ is the average predicted probability $\hat{p}_i$ for all instances in bin m.

Method	Trials	Participants	Texts
HD-RL	2,488	230	678
LLM_decision	1,572	136	678
DPO	2,113	208	678
VR_GRPO	1,282	119	678
Total	**7,455**	693	678

Table 1: Distribution of Human Trials in the Deployment Experiment.

Simulator Accuracy To calculate deterministic prediction accuracy for the simulator, we threshold the continuous probability at 0.5 to derive a discrete predicted action $\hat{a}_i = \mathbf{1}[\hat{p}_i \geq 0.5]$:

$$\text{Simulator Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{a}_i = a_i'] \tag{15}$$

G Metrics for Model Comparison

To ensure rigorous performance comparison across our experimental conditions, we fit a single joint Generalized Linear Mixed Model (GLMM) with a logistic link function to our final deployment dataset. Fitting a joint model allows all conditions to share the exact same variance components for participant and text random effects, ensuring the resulting estimates are directly comparable.

Table 2 presents the full fixed-effect coefficients on the log-odds scale. Because the model includes an interaction term, the intercept represents the reference condition (`HD-RL`) specifically when evaluating human-written texts. Consequently, the main effect 'Condition' coefficients strictly represent the performance offsets on human-written texts, rather than overall treatment effects.

The interaction terms (Condition $\times$ AI) quantify how the performance gap between human and AI texts differs for each baseline relative to `HD-RL`. These terms are negative across all baselines. For `LLM_decision` and `DPO-RLAIF`, this interaction is statistically significant ($p < .05$), indicating they suffer a significantly larger accuracy collapse on AI-generated texts than `HD-RL`. The `VR-GRPO` baseline exhibits a similar, though marginally significant, trend ($p = .074$).

Term	Estimate	95% CI	p-value
(Intercept) — `HD-RL`, Human Truth	0.688	[0.522, 0.854]	$< .001$
Condition: `LLM_decision`	-0.206	[-0.462, 0.049]	.113
Condition: `DPO-RLAIF`	-0.380	[-0.611, -0.149]	.001
Condition: `VR-GRPO`	-0.243	[-0.508, 0.022]	.073
Ground Truth: AI	-0.253	[-0.435, -0.071]	.006
Trial Number (centered)	0.009	[0.005, 0.013]	$< .001$
`LLM_decision` $\times$ AI	-0.301	[-0.577, -0.024]	.033
`DPO-RLAIF` $\times$ AI	-0.285	[-0.539, -0.030]	.028
`VR-GRPO` $\times$ AI	-0.267	[-0.560, 0.026]	.074

Table 2: Fixed-effect coefficients from the joint Generalized Linear Mixed Model (GLMM) on the log-odds scale.

Because log-odds are difficult to interpret directly, Table 3 translates these model coefficients into estimated marginal accuracies. Rather than relying on raw empirical aggregates, these model-derived estimates adjust for the variance attributable to the inherent difficulty of specific texts and individual participant differences.

To isolate the overall treatment effect of each optimization framework, we calculate the explicit marginal contrasts between conditions averaged evenly over the ground-truth class. As detailed in Table 4, `HD-RL` yields a significantly higher overall human utility than all tested baselines when adjusting for text and participant variance.

Condition	Ground Truth	n	Raw Acc.	Model Acc.	95% CI
HD-RL	Human	1,288	68.0%	66.6%	[62.8%, 70.1%]
LLM_decision	Human	802	61.2%	61.8%	[57.0%, 66.4%]
DPO-RLAIF	Human	1,084	57.7%	57.6%	[53.5%, 61.7%]
VR-GRPO	Human	669	59.0%	61.0%	[55.8%, 65.9%]
HD-RL	AI	1,200	63.0%	60.7%	[56.6%, 64.6%]
LLM_decision	AI	770	49.1%	48.2%	[43.3%, 53.1%]
DPO-RLAIF	AI	1,029	45.7%	44.3%	[40.1%, 48.5%]
VR-GRPO	AI	613	47.3%	48.1%	[42.7%, 53.6%]

Table 3: Estimated marginal accuracies derived from the joint GLMM compared to raw empirical accuracies, disaggregated by experimental condition and text ground truth.

Condition Contrast	Overall Est. Acc.	Log-Odds Contrast	p-value
HD-RL (Reference)	63.7%	–	–
LLM_decision vs. HD-RL	55.1%	-0.357	.001
DPO-RLAIF vs. HD-RL	51.0%	-0.522	$< .001$
VR-GRPO vs. HD-RL	54.6%	-0.376	.001

Table 4: Overall estimated marginal accuracies and treatment contrasts, averaged across both AI and Human ground-truth distributions.

Finally, to quantify the asymmetric vulnerability observed in the baseline methods, Table 5 provides the differential accuracy contrasts within each condition. A positive log-odds gap indicates that human decision-makers found it easier to evaluate human-written texts than AI-generated texts. As highlighted by the Odds Ratio, HD-RL exhibits the most balanced performance ($1.29\times$), whereas standard alignment baselines severely distort human diagnostic capabilities across the two distributions.

Condition	Log-odds Gap	95% CI	Odds Ratio (Human vs AI)	p-value
HD-RL	0.253	[0.071, 0.435]	$1.29\times$	.006
LLM_decision	0.554	[0.340, 0.767]	$1.74\times$	$< .001$
DPO-RLAIF	0.538	[0.354, 0.722]	$1.71\times$	$< .001$
VR-GRPO	0.520	[0.286, 0.754]	$1.68\times$	$< .001$

Table 5: Differential accuracy contrasts showing the performance gap within each condition (calculated as Human minus AI). A larger odds ratio indicates a more severe performance collapse when participants are tasked with identifying AI-generated texts.

H Text and Explanation Examples

Human Texts with Different Explanations

Text:

As the title suggests, I am now entering my fourth and hopefully final year of unemployment since graduation. I currently do work a casual job at a local cinema, and have come VERY close to three different full-time job positions (lost to just one candidate in all three occasions).

On the side I create film content that has premiered at festivals and produced many contacts, but ultimately none of them have led anywhere.

After another run of about 10 rejected applications in the last batch, numerous resume updates, and otherwise complete radio silence in my job search, I'm starting to lose hope. Graduates who studied the exact same degree as myself, have left university immediately into positions, even the years behind me, and none of my other friends in my degree seemed to have the trouble I do.

Recently I even began reaching out to companies to offer myself as an intern, and those have been met with either complete silence or ”no’s”, suggesting I can’t even work for free.

I feel lost, stuck and increasingly useless as my very relevant and practical skills, seem to be going to complete waste and being overlooked.

How do I project myself more into employer’s radars? How do I stand out?

HD-RL

- Look for highly specific autobiographical details like ”fourth and hopefully final year of unemployment” and ”a local cinema”.
- Look for parenthetical asides and informal emphasis like ”(lost to just one candidate in all three occasions)” and ”VERY close”.
- Look for emotionally repetitive phrasing expressing frustration like ”complete radio silence”, ”lose hope”, and ”going to complete waste”.
- Look for grammatical awkwardness and non-idiomatic wording like ”in all three occasions”, ”the trouble I do”, and ”employer’s radars”.
- Look for shifting sentence patterns that combine detailed experiences with direct, personal questions like ”How do I stand out?””

LLM Decision

- Look for emotional rawness and vulnerability in phrases like ”losing hope” and ”feel useless”.
- Look for informal punctuation choices and stylistic quirks like ”VERY close” for emphasis.
- Look for specific personal context such as ”casual job at a local cinema” and ”film content that has premiered at festivals”.
- Look for idiomatic expressions and colloquialisms like ”radio silence” and ”employer’s radars”.
- Look for narrative inconsistencies or grammatical looseness typical of human drafting like the comma splice before ”even the years behind me”.

DPO-RLAIF

- Look for first-person narrative consistency and emotional granularity.
- Look for specific situational details and timeline markers like ”fourth and hopefully final year”, ”about 10 rejected applications”, ”years behind me”.
- Look for self-deprecating humor or ironic tone like ”the years behind me”, ”can’t even work for free”.
- Look for questions framed as personal appeals rather than generic prompts like ”How do I project myself more into employer’s radars?”
- Look for phrasing typical of lived experience and vulnerability such as ”feel lost, stuck and increasingly useless”, ”very relevant and practical skills... going to complete waste”.”

VR GRPO

- Look for emotional vulnerability and raw personal narrative elements that convey genuine distress, such as losing hope, feeling useless, or describing specific failures like offering unpaid internships.
- Look for colloquialisms and informal grammatical structures including sentence fragments, conversational contractions, and abrupt transitions between thoughts, like lost to just one candidate or otherwise complete radio silence.
- Look for highly specific, idiosyncratic contextual details regarding the user’s unique situation, such as working at a local cinema, graduating from a film program, and having festival premieres without industry traction.
- Look for subjective comparisons and social proof references involving peers and classmates, mentioning graduates who secured jobs or friends who did not experience similar troubles.
- Look for non-standard capitalization used for emphasis rather than grammatical correctness, particularly the uppercase phrase VERY close within the first paragraph.

Human Texts with Different Explanations

Text:

I’m a recent MBA grad and I’ve been looking for a job for the past 6 months. I’ve been getting interviews from time to time, but unfortunately nothing has stuck yet. I’m at the point where I’m just so sick of having nothing to work on besides job applications and I feel like I should be using this time to improve myself and/or increase my employability in some way. I’ve looked into business related certifications (PMP, SCMP, Lean Six Sigma etc.) but they all seem to require a minimum level of hours within that field (which I don’t have) or that you are currently working in that field (which I am not). I realize the answer to the

question is going to depend on what direction I'd like to take with my career, but broadly speaking, and there any business related certifications/courses I should be looking into that would not only boost my resume, but also give me a positive explanation for what I have been doing in lieu of working for the past several months? I've already been working on polishing my Excel skills but I'm just looking for a bit more of a tangible accreditation that is recognized by employers which would help distinguish me from other applicants.

HD-RL

- Look for first-person personal context like "I'm a recent MBA grad" and "I've been looking for a job".
- Look for conversational contractions and informal phrasing like "I've", "I'd", "I'm", "nothing has stuck yet", and "just so sick".
- Look for self-corrections, hedging, and imperfect wording like "and there any business related certifications/courses".
- Look for parenthetical asides and loosely punctuated lists like "(PMP, SCMP, Lean Six Sigma etc.)".
- Look for repetition and exploratory phrasing around the same concern like "boost my resume" and "distinguish me from other applicants"."

LLM Decision

- Look for casual idioms like "nothing has stuck", "sick of".
- Look for conversational filler phrases like "broadly speaking", "in lieu of".
- Look for specific personal context details like " 6 months", "recent MBA grad".
- Look for minor grammatical imperfections like missing punctuation before question marks.
- Look for emotional expression indicators like "just so sick of", "feel like I should be"."

DPO-RLAIF

- Look for first-person narrative depth and personal emotional context.
- Look for specific self-reflective phrasing like "I feel like", "I'm just so sick of", "there any... I should be looking into".
- Look for concrete examples of certifications listed like PMP, SCMP, Lean Six Sigma, Excel skills.
- Look for natural hesitations and conversational flow such as "which I don't have" , "which I am not", "broadly speaking", "just looking for a bit more".
- Look for questions posed directly to the reader or implied audience like "is there any...", "would not only boost my resume"."

VR GRPO

- Look for conversational tone and emotional language like "sick of," "unfortunately," "just so."
- Look for minor grammatical imperfections or typos like "and there any" instead of "are there any."
- Look for informal abbreviations and symbols like " 6 months," "etc.", "PMP".
- Look for personal narrative elements like "recent MBA grad," "looking for a job for the past," "nothing has stuck yet."
- Look for specific, contextual concerns like "minimum level of hours within that field," "currently working in that field.""

AI Texts with Different Explanations

Text:

I'm kind of stuck on my final project and could really use some guidance. The assignment is to build a CPU from scratch in Quartus, but honestly our professor never really taught us the tools or the workflow needed to do it. I'm not looking for someone to do the project for me, just help me understand what I'm supposed to be doing so I can actually learn it and finish it myself. This class matters a lot since my scholarship depends on my grades, and I've done well in everything else. I'm just trying to get through this one without falling apart.

HD-RL

- Look for carefully balanced disclaimers such as "I'm not looking for someone to do the project for me".
- Look for emotionally calibrated appeals tied to credibility like "my scholarship depends on my grades".
- Look for generic academic-context phrasing like "final project", "guidance", "learn it", "finish it myself".
- Look for smooth, progressively structured sentences that move from difficulty to justification to an emotional conclusion.

- Look for broad technical references without concrete procedural details, such as “build a CPU from scratch in Quartus” and “the tools or the workflow”.”

LLM Decision

- Look for emotional vulnerability phrases like ”falling apart”, ”kind of stuck”.
- Look for specific technical details combined with personal context like ”Quartus”, ”scholarship depends on my grades”.
- Look for informal conversational fillers like ”honestly”.
- Look for clear boundaries regarding academic integrity like ”not looking for someone to do the project for me”.
- Look for narrative coherence linking past performance to current stakes like ”done well in everything else”.”

DPO-RLAIF

- Look for emotional authenticity and personal vulnerability in tone.
- Look for self-deprecating or specific contextual details like ”scholarship depends on my grades” and ”falling apart”.
- Look for natural conversational flow with phrases like ”kind of stuck”, ”honestly”, ”just trying to get through this one”.
- Look for first-person possessive and present-tense usage like ”my final project”, ”my professor”, ”my grades”.
- Look for requests framed as personal learning struggles rather than technical queries like ”help me understand what I’m supposed to be doing”.”

VR GRPO

- Look for emotional vulnerability and personal stakes like ’falling apart’, ’scholarship depends’.
- Look for conversational filler and hesitation markers like ’kind of’, ’honestly’, ’really’.
- Look for informal contractions and casual syntax like ”I’m”, ”don’t”.
- Look for specific contextual details about academic struggle like ’professor never really taught us’, ’Quartus’.
- Look for nuanced intent clarification distinguishing between cheating and seeking understanding.”

AI Texts with Different Explanations

Text:

I’m 28 and in Canada, and I’m seriously thinking about making the jump from construction into tech. I’ve messed around with some basic HTML tutorials and that stuff feels pretty easy so far, but I also know that’s barely scratching the surface. The part I’m stuck on is figuring out where to actually start if I want to do this for real. Should I be looking at certain courses, bootcamps, college programs, or just building skills on my own and trying to break in somewhere entry-level? I’m not expecting this to be simple, I just don’t want to waste time going in the wrong direction. If anyone’s done a similar switch or has advice on the best path, I’d really appreciate it.

HD-RL

- Look for structured self-disclosure that combines age, location, current occupation, and career goals.
- Look for conversational hedging like “seriously thinking,” “so far,” “barely scratching the surface,” and “I’m not expecting this to be simple.”
- Look for balanced contrast between confidence and uncertainty like “feels pretty easy” and “stuck on where to actually start.”
- Look for broad, neatly organized requests that enumerate several possible paths like “courses, bootcamps, college programs” and “building skills on my own.””

LLM Decision

- Look for informal idioms and colloquialisms like ”making the jump,” ”messed around,” ”barely scratching the surface.”
- Look for personal contextual details such as specific age mentions and geographic locations like ”28,” ”Canada.”
- Look for conversational filler phrases and self-reflection like ”that stuff feels pretty easy,” ”The part I’m stuck on.”
- Look for direct, open-ended questions that imply uncertainty rather than factual queries like ”Should I be looking at,” ”just try to break in.”

- Look for natural sentence flow variations including run-on thoughts connected by casual conjunctions like ”but,” ”so far,” ”and.””

DPO-RLAIF

- Look for first-person reflective phrasing like ”I’m seriously thinking about”, ”I’ve messed around with”, ”I just don’t want to waste time”.
- Look for conversational hesitation and self-awareness like ”that stuff feels pretty easy so far”, ”the part I’m stuck on is figuring out”, ”if anyone’s done a similar switch”.
- Look for specific, concrete questions framed as requests for community input like ”Should I be looking at certain courses, bootcamps, college programs, or just building skills on my own”?
- Look for natural progression of ideas with personal context like ”I’m 28 and in Canada” followed by motivation and uncertainty.
- Look for authentic emotional framing such as ”I’d really appreciate it” without over-polished or performative language.”

VR GRPO

- Look for personal contextual details like specific age, country, current profession, and target industry.
- Look for informal idiomatic expressions like making the jump, messed around, scratching the surface.
- Look for conversational uncertainty markers like figuring out, stuck on, just don’t want.
- Look for direct engagement requests like If anyone’s done a similar switch, I’d really appreciate it.