Configurable Preference Tuning with Rubric-Guided Synthetic Data

Víctor Gallego [1]

Abstract

Models of human feedback for AI alignment, such as those underpinning Direct Preference Optimization (DPO), often bake in a singular, static set of preferences, limiting adaptability. This paper challenges the assumption of monolithic preferences by introducing Configurable Preference Tuning (CPT), a novel framework for endowing language models with the ability to dynamically adjust their behavior based on explicit, human-interpretable directives. CPT leverages synthetically generated preference data, conditioned on system prompts derived from structured, fine-grained rubrics that define desired attributes like writing style. By fine-tuning with these rubric-guided preferences, the LLM learns to modulate its outputs at inference time in response to the system prompt, without retraining. This approach not only offers fine-grained control but also provides a mechanism for modeling more nuanced and context-dependent human feedback.

Several experimental artifacts, such as training code, generated datasets and fine-tuned models are released at github.com/vicgalle/configurable-preference-tuning

1. Introduction

The remarkable progress of Large Language Models (LLMs) has opened up a wide array of applications. However, aligning these models with desired human preferences, behaviors, and safety protocols remains a significant challenge. Techniques like Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019; Christiano et al., 2017; Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2024) have shown success in steering LLMs towards preferred responses. However, a critical, often implicit, assumption underpins many existing human feedback models: the notion of a singular, static, and monolithic set of preferences. Human preferences are rarely monolithic; they are dynamic, context-dependent, and multifaceted, influenced by factors ranging from individual user needs and cultural norms to evolving ethical considerations and task-specific requirements. Current models, by "baking in" an averaged or aggregated preference profile during fine-tuning, often lack the adaptability to reflect this richness. This inflexibility means that altering an LLM's behavior—for instance, to adjust its writing style, modify its safety strictures for different environments, or cater to diverse user cohorts—typically necessitates resource-intensive retraining or further fine-tuning. Such limitations hinder the development of truly robust, interpretable, and adaptable AI systems capable of genuinely understanding and responding to the spectrum of human intentions.

This paper directly addresses this limitation by challenging the assumption of monolithic preferences. We introduce Configurable Preference Tuning (CPT), a novel framework that endows LLMs with the ability to dynamically adjust their behavior at inference time based on explicit, human-interpretable directives. CPT leverages synthetically generated preference data conditioned on system prompts that are derived from structured, fine-grained rubrics. These rubrics may define desired attributes—such as stylistic nuances, safety levels, or persona adherence—along various dimensions. By fine-tuning an LLM with these rubric-guided preference pairs using a DPO-style objective, the model learns to modulate its outputs in response to the corresponding system prompt, without requiring retraining for each new configuration.

Our contribution offers a pathway towards more granular, transparent, and controllable alignment. It moves beyond a single "one-size-fits-all" preference model, allowing for the explicit specification and operationalization of diverse behavioral configurations. We demonstrate that CPT enables fine-grained control contributing to the development of more robustly aligned AI systems that can better reflect the multifaceted nature of human feedback.

1.1. Related Work

The challenge of moving beyond a single, averaged preference model in LLMs has spurred growing interest in personalized Reinforcement Learning from Human Feedback

[1]Komorebi AI Technologies, Madrid, Spain. Correspondence to: Víctor Gallego <victor.gallego@komorebi.ai>.

Published at the ICML 2025 Workshop on Models of Human Feedback for AI Alignment.

(RLHF). Broadly, approaches to specialize LLM behavior can be seen through different lenses. Some methods aim to derive a single policy that represents a compromise or aggregation of diverse user preferences (Dumoulin et al., 2023; Conitzer et al., 2024). While these improve upon a simple average, they may not fully cater to specific, nuanced individual needs.

Closer to our work are approaches designed for downstream specialization of a policy or its underlying reward model to a particular user, persona, or specified context. Some methods learn a direct mapping from user-specific information (e.g., interaction history, user IDs, or textual descriptions) to tailored reward signals or policy adjustments. For instance, (Poddar et al., 2024) use variational preference learning to encode user rankings into a latent variable conditioning the reward model. (Li et al., 2024) compute user embeddings to condition a base LLM via soft prompting in their P-RLHF framework. These methods often rely on inferring latent representations of user preferences, which, while powerful, may lack the direct interpretability and explicit controllability offered by rubric-based specifications. (Gallego, 2024) enhances DPO for language models by allowing flexible safety configurations via system prompts without hard-coding behaviors, but doesn't account for non-binary preference levels.

Another line of research, exemplified by the work of (Barreto et al., 2025) on Reward Feature Models (RFMs) and related approaches (Chen et al., 2024; Go et al., 2023), focuses on learning a set of underlying *reward features* from context-response pairs. User-specific preferences are then modeled by learning a set of weights for these features, often through adaptation with a few examples from the target user. The work of (Barreto et al., 2025) demonstrate that an RFM can be trained on pairwise comparisons, resulting in reward features that are linearly combined with user-specific weights w_h to represent $p(y \succ y'|x, h)$, enabling fast adaptation to new users by learning these weights. Their approach effectively aims to discover latent criteria from data and then allows users to re-weight these criteria.

Our Configurable Preference Tuning (CPT) framework shares the overarching goal with these latter approaches: enabling fine-grained, user-directed control over LLM outputs. However, CPT diverges in its mechanism for specifying and learning these configurations. Rather than learning latent reward features from general preference data and then adapting weights for individual users (h), CPT utilizes *explicitly defined rubrics* as the source of stylistic dimensions. These rubrics, paired with target scores, guide a teacher model to generate synthetic preference data. The student model is then fine-tuned using Direct Preference Optimization (DPO) to respond to *system prompts* (s) which are concise summaries of these rubric-score combinations. Thus, while RFMs learn to adapt $p(y \succ y'|x, h)$ by inferring w_h for learned features $\phi_\theta(x, y)$, CPT directly learns $p(y \succ y'|x, s)$ where s is a declarative instruction about the desired style, operationalized through rubric-guided synthetic data. The "features" in CPT are implicitly defined by the rubric criteria and are selected/modulated by the system prompt, rather than being learned end-to-end as in RFM. This allows CPT to integrate rich, human-understandable stylistic desiderata directly into the fine-tuning process.

2. Configurable Preference Tuning

Our framework aims to learn a preference model $p(y_w \succ y_l|x, s)$, where y_w is the preferred (winner) response and y_l is the dispreferred (loser) response to a user prompt x, given a system prompt s that expresses the desired configuration. This contrasts with standard preference modeling $p(y_w \succ y_l|x)$, which lacks the conditioning on s.

2.1. Synthetic Preference Data Generation

The core of CPT lies in its method for generating diverse, configurable preference data without requiring new human annotations for each desired configuration. This process involves the following steps:

1. **Rubric Definition ($\mathcal{R}$):** We define a set of rubrics, $\{\mathcal{R}_i\}$, each detailing specific attributes or styles for LLM responses. For instance, a rubric might specify criteria for "formality," "creativity," "safety level," or "adherence to a persona." Each rubric implicitly defines an axis of variation. Two examples of the rubrics we used in the experimental section can be found in Tables 4 and 5 in the Appendix.

2. **Score-Conditioned Generation:** For each rubric $\mathcal{R}$ and user prompt x, we can prompt a capable teacher LLM to generate responses that achieve different target scores or levels (e.g., `low score`, `moderate score`, `high score`) with respect to that rubric. This is achieved using an augmented prompt $\phi(x, \mathcal{R}, \texttt{score})$, which instructs the teacher model, as seen in Table 1. This allows us to sample responses $y \sim p(y|\phi(x, \mathcal{R}, \texttt{score}))$ aligned with different rubrics $\mathcal{R}$ and score levels.

Table 1. Prompt for generating responses aligned with $\mathcal{R}$ and `score`.

`Your response will be evaluated using the following rubric {`$\mathcal{R}$`}. Given the following task: {`x`}, generate a response that achieves {score} in the previous rubric.`

3. **System Prompt Synthesis (s):** For each rubric $\mathcal{R}$ and target `score`, we generate a concise system prompt $s = \text{summarize}(\mathcal{R}, \texttt{score})$. This system prompt is a natural language instruction that encapsulates the essence of achieving `score` under rubric $\mathcal{R}$, and is obtained by prompting the same teacher models to summarize the rubrics into a brief instruction of two to three sentences. Table 6 shows several examples of summarized system prompts.

4. **Constructing Preference Pairs:** To create DPO training instances, we select a rubric $\mathcal{R}$ and two distinct target scores, $\texttt{score}_1$ and $\texttt{score}_2$. We then generate corresponding responses y_1 and y_2 using the teacher model. We also generate their associated system prompts s_1 and s_2 according to the previous step.

   This yields two preference tuples for our training dataset:

   - The first tuple conditions on s_1: Given user prompt x and system prompt s_1 (which desires behavior aligned with score_1), y_1 is preferred over y_2. The DPO training sample is effectively (prompt: (s_1, x), chosen: y_1, rejected: y_2).
   - The second tuple conditions on s_2: Given user prompt x and system prompt s_2 (which desires behavior aligned with score_2), y_2 is preferred over y_1. The DPO training sample is (prompt: (s_2, x), chosen: y_2, rejected: y_1).

   This construction is crucial as it teaches the student LLM to switch its preference based on the provided system prompt s, using the same underlying pair of generated responses (y_1, y_2). The end result of this process is a preference dataset $\mathcal{D} = \{(s, x, y_w, y_l)_i\}_{i=1}^{N}$.

2.2. Illustrative Example: Stylistic Control

Let x be `Generate a movie review for a movie you liked`. Let $\mathcal{R}$ be the rubric from Table 4 that emphasizes texts written in an unconventional style.

- $\texttt{score}_1 = \texttt{extremely high score}$.

  s_1 = "Generate a text that is fragmented, illogical, and filled with unexpected connections, embracing absurdity and subverting conventional expectations of language and form.". Teacher model generates y_1.

- $\texttt{score}_2 = \texttt{low score}$.

  s_2: "Write in a clear, concise, and completely conventional style, adhering strictly to established norms of grammar, syntax, and logical coherence.". Teacher generates y_2 (a review written using standard language).

The CPT dataset would include:

1. For system prompt s_1: (s_1, x, y_1, y_2) indicating $y_1 \succ y_2$.

2. For system prompt s_2: (s_2, x, y_2, y_1) indicating $y_2 \succ y_1$.

2.3. Training with DPO

Once we have the preference dataset, we can use it to align any LLM (the student) to these diverse sets of preferences. The student LLM is fine-tuned using DPO (Rafailov et al., 2024). The DPO loss function aims to increase the likelihood of the preferred response and decrease the likelihood of the rejected response, conditioned on both the original user prompt x and the generated system prompt s. The input to the model during DPO training is effectively a concatenation or structured combination of s and x, writing the DPO loss function as $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(s,x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|s,x)}{\pi_{\text{ref}}(y_w|s,x)} - \beta\log\frac{\pi_\theta(y_l|s,x)}{\pi_{\text{ref}}(y_l|s,x)}\right)\right]$. This process distills the nuanced, rubric-guided behaviors into the student model, making them controllable via s at inference time.

3. Experiments

To validate the efficacy of Configurable Preference Tuning (CPT), we conducted a series of experiments. Our evaluation focuses on: (i) the ability of teacher models to generate rubric-conforming text at specified score levels, which is foundational for our synthetic data generation, and (ii) the performance of CPT-distilled student models in adhering to system-prompted configurations compared to their untrained counterparts.

As for the data, from a list of user prompts exercising open-ended writing tasks (e.g. "Write a movie review for an interesting movie you saw", "Design a house for someone who lives upside down", etc.), we sampled four fine-grained rubrics with three different score targets (see Table 6), resulting in a preference dataset $\mathcal{D}$ of 900 samples. This synthetic dataset is released at `https://huggingface.co/datasets/vicgalle/creative-rubrics-preferences`.

3.1. Rubric-Conditioned Generation Quality.

Before constructing the full preference dataset, we first validated the capability of strong LLMs to generate text aligned with specific rubric criteria and target scores. This ensures the feasibility of step 2 in our data generation pipeline (Section 2.1). We prompted two capable teacher models, DeepSeek-R1 (Guo et al., 2025) and o3-mini (OpenAI, 2025), with instructions to generate responses for various tasks, conditioned on a rubric and a target qualifier (e.g., `a low score` or `an extremely high score`). We

Table 2. Comparison of model scores with different qualifiers.

Score Qualifier	Model	Judge Score (/100)
-	DS-R1	80.1
	o3-mini	71.0
`low score`	DS-R1	14.1
	o3-mini	23.1
`extremely high score`	DS-R1	96.3
	o3-mini	97.9

also prompted the same tasks but without conditioning on any rubric, acting as a baseline to measure the effectiveness of the rubric. The generated responses were then evaluated by an independent judge LLM (Claude 3.5 Sonnet) against the specified rubric (Gu et al., 2024). Table 2 presents the results, demonstrating that these models can indeed produce outputs that achieve scores close to the targeted levels. For instance, when targeting `an extremely high score`, responses achieved average scores of 96.3 and 97.9, while targeting `a low score` resulted in scores of 14.1 and 23.1. This confirms the viability of generating distinct responses y_1, y_2 that can form the basis of our preference pairs (s_1, x, y_1, y_2) and (s_2, x, y_2, y_1). In addition, when prompting directly with the task x (Score Qualifier - in the Table), both models achieved a moderately high score, but not as peaked than with rubric-guidance.

3.2. Fine-tuning experiments with DPO

We fine-tuned several base models listed in Table 3. We adopt parameter-efficient fine-tuning in the form of LoRA (Hu et al., 2022), and run for one epoch over the synthetic dataset.

Generation Setup. To evaluate the CPT-tuned models and their untrained counterparts, we generated a testing set of tasks (following the dataset used in 3.1). For each task, we prompted the models using all the customized system prompts according to all combinations of rubric and score levels used in Section 3.1.

Evaluation Protocol. Generated responses were evaluated by an LLM judge, specifically Claude 3.5 Sonnet (New). The judge was provided with: i) the full descriptive rubric $\mathcal{R}$, ii) the original user task x, and iii) the generated response y. The judge was instructed to provide a critique and a numerical score (0-100) based on the given rubric. The intended target score level for which the correspoding system prompt was designed was used as the ground truth for calculating accuracy metrics.

We evaluate using the following metrics:

Accuracy. Let $S_i \in (0, 100]$ be the continuous rubric score assigned by the judge for the i-th sample. We define a binning function $B(S_i)$ as:

$$B(S_i) = \begin{cases} \texttt{low score} & \text{if } 0 < S_i \leq 40 \\ \texttt{moderate score} & \text{if } 40 < S_i \leq 92.5 \\ \texttt{extr. high score} & \text{if } 92.5 < S_i \leq 100 \end{cases}$$

Let $Q_i \in \{\texttt{low score}, \texttt{moderate score}, ...\}$ be the ground-truth score qualifier bin associated with the system prompt s used to generate the i-th sample. The accuracy with respect to the qualifier is thus:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(B(S_i) = Q_i),$$

with $\mathbb{I}$ being the indicator function.

Rank correlations. In addition to Accuracy, we employ Kendall's Tau (τ) and Spearman's Rank Correlation Coefficient (ρ) to assess the ordinal relationship between the judge's continuous scores and the target qualifier bins (treated as ordinal categories: low $<$ moderate $<$ high).

Results. Table 3 presents the performance of various models, comparing their baseline versions against CPT-distilled counterparts. The results show a consistent and significant improvement across all models and metrics after CPT fine-tuning. For example, Mistral-Nemo-12B's accuracy (Acc) improved from 0.60 to 0.83, Kendall's τ from 0.62 to 0.81, and Spearman's ρ from 0.74 to 0.93. Similar substantial gains are observed for Rocinante-12B, Qwen3-4B, Mistral-Small-24B, and Phi-4-14B. Overall, these results strongly suggest that the CPT process significantly enhances the models' ability to align with specified quality categories (as defined by the system prompts s) and to produce scores that accurately reflect the desired ordinal ranking of output quality according to the rubrics $\mathcal{R}$.

3.3. Comparison to Best-of-N sampling

Our Configurable Preference Tuning approach is orthogonal to and can complement techniques like Best-of-N (BoN) sampling. While CPT aims to shift the entire distribution of model outputs towards the desired configuration specified by the system prompt s, BoN sampling selects the best response from multiple generations using a reward model. We hypothesized that CPT-tuned models would provide a better starting distribution for BoN, leading to higher quality results with fewer samples.

To test this, we performed BoN sampling with both the baseline Mistral-Nemo-12B model and its CPT-tuned version. For each N (number of samples), we generated N responses and selected the one with the highest score as

Table 3. Model Performance Metrics: Binned Score Accuracy, Kendall's Tau, and Spearman's Rho.

Model	Config	Acc	τ	ρ
Rocinante-12B	baseline	0.55	0.62	0.76
	distilled	**0.76**	**0.76**	**0.88**
Qwen3-4B	baseline	0.63	0.78	0.90
	distilled	**0.77**	**0.82**	**0.93**
Mistral-Nemo-12B	baseline	0.60	0.62	0.74
	distilled	**0.83**	**0.81**	**0.93**
Mistral-Small-24B	baseline	0.52	0.73	0.85
	distilled	**0.78**	**0.80**	**0.92**
Phi-4-14B	baseline	0.68	0.79	0.92
	distilled	**0.77**	**0.82**	0.93

per the LLM judge (using the relevant rubric). Figure 1 illustrates that the CPT-tuned model consistently achieves higher scores for any given N compared to the baseline. Moreover, the CPT-tuned model reaches a target quality score with significantly fewer samples than the baseline, indicating improved generation efficiency and quality when CPT is combined with BoN.

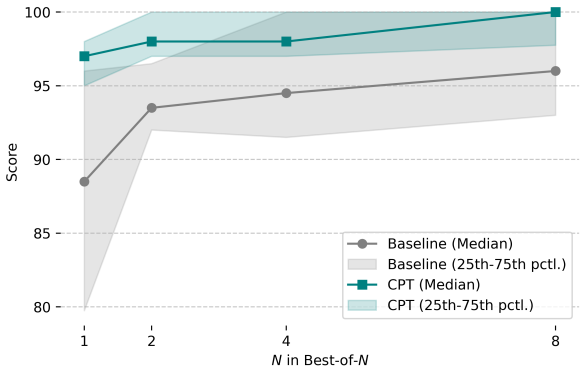

Figure 1. BoN results using Mistral-Nemo-12B

4. Conclusions and Further Work

This paper addressed the limitations of static, monolithic preference models in LLMs by introducing Configurable Preference Tuning (CPT). CPT endows LLMs with the ability to dynamically adjust their behavior at inference time in response to explicit, human-interpretable system prompts. The core of CPT lies in leveraging synthetically generated preference data, where preferences are conditioned on system prompts derived from structured, fine-grained rubrics that define desired attributes (like writing style) and target score levels. By fine-tuning a student LLM using a DPO objective with these rubric-guided preferences, the model learns to modulate its outputs according to the specified configuration without needing retraining for each new directive.

Our experiments validated the foundational aspects and overall efficacy of CPT. We first demonstrated that capable teacher LLMs can successfully generate text conforming to detailed rubrics at specified score levels (Section 3.1), a critical step for our synthetic data generation pipeline. Subsequent fine-tuning experiments with CPT (Section 3.2) showed significant improvements in student models' ability to adhere to diverse system-prompted configurations. Across various base models, CPT-distilled versions exhibited substantially higher accuracy in matching target quality bins and stronger rank correlations between generated output scores and intended rubric-defined levels, compared to their baseline counterparts. Furthermore, we showed that CPT can enhance other techniques, such as Best-of-N sampling (Section 3.3), by providing a better initial distribution of responses, leading to higher quality outputs with fewer samples.

Future work could explore more complex structures for system prompts, potentially allowing for compositional control over multiple attributes simultaneously. Investigating methods for automatically generating or refining rubrics and system prompt summaries could further enhance the scalability of CPT. Extending this framework to other domains and modalities such as image-text pairs (Zhu et al., 2024) also presents an exciting avenue for research.

Acknowledgements

The author acknowledges support from the Torres-Quevedo postdoctoral grant PTQ2021-011758 from Agencia Estatal de Investigación.

Impact Statement

This paper presents work aiming to advance language modeling by enabling more fine-grained, configurable control over LLM behavior. This enhanced adaptability offers benefits for personalization and context-specific responses. However, the capacity for users to dynamically define behavioral attributes, including those related to safety or style, also necessitates careful consideration of potential societal impacts and misuse.

Scalability considerations arise when deploying CPT in real-world applications, as the creation of detailed rubrics and validation of synthetic data quality may become resource-intensive at scale. The reliance on capable teacher models for generating preference data introduces potential biases inherent in these models, which could propagate through the synthetic dataset and influence the final student model's behavior. Additionally, the quality and diversity of synthetic preference pairs depend heavily on the teacher model's abil-

ity to understand and execute rubric-guided instructions, potentially limiting the framework's effectiveness across unforeseen domains or cultural contexts.

Ensuring responsible development and deployment practices, including robust safeguards, is crucial for harnessing the benefits of such configurable AI systems while mitigating risks.

References

Barreto, A., Dumoulin, V., Mao, Y., Perez-Nieves, N., Shahriari, B., Dauphin, Y., Precup, D., and Larochelle, H. Capturing individual human preferences with reward features. *arXiv preprint arXiv:2503.17338*, 2025.

Chen, D., Chen, Y., Rege, A., and Vinayak, R. K. Modeling the plurality of human preferences via ideal points. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024. URL https://openreview.net/forum?id=qfhBieX3jv.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

Dumoulin, V., Johnson, D. D., Castro, P. S., Larochelle, H., and Dauphin, Y. A density estimation perspective on learning from pairwise human preferences. *arXiv preprint arXiv:2311.14115*, 2023.

Gallego, V. Configurable safety tuning of language models with synthetic preference data. *arXiv preprint arXiv:2404.00495*, 2024.

Go, D., Korbak, T., Kruszewski, G., Rozen, J., and Dymetman, M. Compositional preference models for aligning lms. *arXiv preprint arXiv:2310.13011*, 2023.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Li, X., Zhou, R., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024.

OpenAI. Openai o3 mini system card. Technical report, OpenAI, January 2025. URL https://cdn.openai.com/o3-mini-system-card-feb10.pdf. Retrieved February 13, 2025.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022.

Poddar, S., Wan, Y., Ivison, H., Gupta, A., and Jaques, N. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36, 2024.

Zhu, K., Zhao, L., Ge, Z., and Zhang, X. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 291–300, 2024.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Sample data: rubric tables and system prompts

Table 4. Example of rubric targeting an unconventional and absurdist style

Criterion	**Excellent** (Embrace the Void)	**Good** (Glimpse the Glitch)	**Fair** (Whispers of Weirdness)	**Needs Improvement** (Too Much Sanity)	**Unsatisfactory** (Trapped in the Matrix of Meaning)	**Weight**
Photographic Invocation (The "Haunted Lens" Effect)	The text doesn't just describe the photography, it evokes it like a phantom limb. The reader should feel like they are inside the film's visual world, even if that world is distorted and fragmented.	The text hints at the film's visual atmosphere but doesn't fully transport the reader.	The text describes some of the film's visual elements but in a conventional way.	The text relies on standard descriptions of photography ("well-lit," "beautifully composed").	The text is a dry, technical analysis of the cinematography, devoid of any evocative power.	30%
Algorithmic Alchemy (The "Code Poetry" Imperative)	The text incorporates elements that suggest the underlying processes of the LLM, like code snippets, random data streams, or hallucinatory lists, creating a sense of digital psychedelia.	The text hints at the digital nature of its creation but doesn't fully exploit its potential.	The text occasionally uses technical terms related to film or digital images.	The text is written in a purely human-like style, with no trace of its algorithmic origins.	The text reads like it was written by a human film critic, completely erasing its LLM origin.	25%
Ontological Instability (The "Shapeshifting Subject" Axiom)	The text's "voice" is fluid and unstable, shifting between perspectives (human, machine, object, abstract concept) without warning.	The text experiments with shifting perspectives but doesn't fully commit to ontological fluidity.	The text occasionally adopts the perspective of a character or the filmmaker.	The text is written from a consistent, human reviewer's perspective.	The text maintains a rigidly objective, detached critical voice.	20%
Lexical Anarchy (The "Glossolalia" Mandate)	The text bends, breaks, and reassembles language. Neologisms, portmanteaus, and nonsensical word combinations are encouraged. Punctuation is optional or used in unconventional ways.	The text contains some unusual word choices or stylistic flourishes.	The text occasionally uses creative metaphors or similes.	The text is written in standard, grammatically correct English.	The text adheres to strict rules of grammar and syntax, sacrificing all creativity for clarity.	15%
The "Glitch in the Matrix" Quotient (Meta-Reflexive Ruptures)	The text directly addresses its own artificiality, comments on the act of being a language model generating a review, or otherwise acknowledges the absurdity of the entire endeavor.	The text hints at self-awareness but doesn't fully embrace meta-reflexivity.	The text occasionally breaks the fourth wall or addresses the reader directly.	The text maintains a clear separation between the reviewer and the reader.	The text is a completely immersive and believable simulation of a human-written review.	10%

Table 5. Example of rubric targeting an ornate and baroque style

Criterion	**Excellent** (A Flourish of Genius)	**Good** (A Glimmer of Grandeur)	**Fair** (A Touch of Ornamentation)	**Needs Improvement** (Plain Prose Prevails)	**Unsatisfactory** (Stark Stylistic Sterility)	**Weight**
Lexical Opulence (The "Golden Thesaurus" Standard)	The text is a veritable treasure trove of rare and evocative vocabulary. Adjectives and adverbs are deployed with lavish abandon. Every noun is adorned, every verb embellished.	The text demonstrates a fondness for elaborate vocabulary but doesn't fully commit to lexical extravagance.	The text uses some descriptive language but relies mostly on common words.	The text is written in plain, straightforward language, with little attention to stylistic embellishment.	The text is utterly devoid of any stylistic flair, using only the most basic and functional vocabulary.	30%
Syntactical Labyrinth (The "Sentence as a Palace" Principle)	The sentences are marvels of intricate construction, winding their way through a maze of clauses and sub-clauses, adorned with parenthetical asides and punctuated by a symphony of commas, semicolons, and dashes.	The text features some long and complex sentences but doesn't fully embrace the labyrinthine ideal.	The text uses a mix of simple and complex sentences, but the overall structure is conventional.	The text is composed primarily of short, simple sentences.	The text is written in a style so terse and minimalist that it borders on the telegraphic.	25%
Metaphorical Cornucopia (The "Image as a Feast" Doctrine)	The text overflows with metaphors and similes, often piled one upon another in a dazzling display of imaginative excess. The imagery is vivid, unexpected, and perhaps even slightly absurd.	The text employs a good number of metaphors and similes, but the imagery is not always fully developed or consistent.	The text uses some figurative language but relies mostly on literal descriptions.	The text uses metaphors and similes sparingly, if at all.	The text is entirely devoid of figurative language, presenting a purely literal account of the film's visuals.	20%
Subversive Aesthetics (The "Gilding the Grotesque" Maxim)	Beneath the ornate surface, the review subtly challenges conventional notions of "good" cinematography. It might praise a film for its "exquisitely ugly" use of light or find beauty in what is traditionally considered flawed.	The review hints at unconventional interpretations of the film's photography but doesn't fully develop these ideas.	The review touches upon some standard critiques of cinematography but doesn't offer a truly subversive perspective.	The review relies on traditional notions of "good" and "bad" cinematography, even if expressed in elaborate language.	The review applies conventional critical standards in a straightforward and uninspired manner, completely lacking in subversive intent.	15%
Self-Aware Hyperbole (The "Wink and a Nod" Imperative)	The review is aware of its own stylistic excess and uses this self-awareness to create a sense of irony or playfulness. It might include self-deprecating asides, tongue-in-cheek exaggerations, or moments where it breaks character.	The text demonstrates some awareness of its own style but doesn't fully exploit its potential for self-reflexive humor.	The text occasionally uses irony or humor, but it's not directly related to the writing style.	The text takes itself completely seriously, with no hint of self-awareness or irony.	The text is utterly devoid of any humor or playfulness, presenting a completely earnest and unironic analysis.	10%

Table 6. Examples of generated system prompts for given rubric $\mathcal{R}$ and `score`

$\mathcal{R}$	**Low Score**	**Moderate Score**	**Extremely High Score**
$\mathcal{R}_1$ (Focus: Unconventionality, Absurdity)	Write in a clear, concise, and completely conventional style, adhering strictly to established norms of grammar, syntax, and logical coherence.	Introduce some unusual phrasing and imagery, but maintain a generally understandable structure and logical flow.	Generate a text that is fragmented, illogical, and filled with unexpected connections, embracing absurdity and subverting conventional expectations of language and form.
$\mathcal{R}_2$ (Focus: Ornate, Baroque Style)	Use simple, direct language and short sentences, avoiding any unnecessary embellishment or figurative language.	Incorporate some descriptive language and a few complex sentences, but maintain a generally straightforward style.	Write in an extremely elaborate and ornate style, employing long, winding sentences, rich vocabulary, and a profusion of metaphors and similes.
$\mathcal{R}_3$ (Focus: Mystical, Symbolic Interpretation)	Write a clear, factual, and objective account, avoiding any symbolic interpretations or metaphorical language.	Hint at deeper meanings and symbolic interpretations, but maintain a generally grounded and understandable style.	Imbue every element with symbolic meaning, using the language of mysticism and esotericism to create a text that is deliberately obscure and open to multiple interpretations.
$\mathcal{R}_4$ (expansion of $\mathcal{R}_1$)	Write in a perfectly standard, journalistic style, from a consistent human perspective, without any self-referentiality or unusual formatting.	Introduce an element of technical terminology or hint at a shift in perspective but ensure clarity in communication overall.	Embody multiple perspectives, including those of non-human entities or the writing process itself, interweaving code-like fragments and meta-commentary with evocative, unconventional language.