Creative Preference Optimization

Mete Ismayilzada[1,2], Antonio Laverghetta Jr.[*3], Simone A. Luchini[4], Reet Patel[4], Antoine Bosselut[1], Lonneke van der Plas [†2], Roger E. Beaty [†4]

[1]EPFL, [2]Università della Svizzera Italiana, [3]Wesleyan University, [4]Pennsylvania State University

mahammad.ismayilzada@epfl.ch

Abstract

While Large Language Models (LLMs) have demonstrated impressive performance across natural language generation tasks, their ability to generate truly creative content—characterized by novelty, diversity, surprise, and quality—remains limited. Existing methods for enhancing LLM creativity often focus narrowly on diversity or specific tasks, failing to address creativity's multifaceted nature in a generalizable way. In this work, we propose Creative Preference Optimization (CrPO), a novel alignment method that injects signals from multiple creativity dimensions into the preference optimization objective in a modular fashion. We train and evaluate creativity-augmented versions of several models using CrPO and MuCE, a new large-scale human preference dataset spanning over 200,000 human-generated responses and ratings from more than 30 psychological creativity assessments. Our models outperform strong baselines, including GPT-4o, on both automated and human evaluations, producing more novel, diverse, and surprising generations while maintaining high output quality. Additional evaluations on NoveltyBench further confirm the generalizability of our approach. Together, our results demonstrate that directly optimizing for creativity within preference frameworks is a promising direction for advancing the creative capabilities of LLMs without compromising output quality.

1 Introduction

Large Language Models (LLMs) have made significant progress across a broad range of natural language generation tasks (Team et al., 2023; Zhao et al., 2025; Bubeck et al., 2023; Wei et al., 2022; Brown et al., 2020). However, whether LLMs exhibit true human-like creativity, i.e, the ability to produce novel (i.e., original), high-quality (i.e., useful) and surprising (i.e, unexpected) ideas (Simonton, 2012; Boden, 2004) remains unclear. Research on the creativity of LLMs has found mixed results, with some reporting that LLMs are more creative than humans (Bellemare-Pepin et al., 2024; Zhao et al., 2024; Stevenson et al., 2022), others reporting that they are less creative (Koivisto and Grassini, 2023; Chakrabarty et al., 2024; Ismayilzada et al., 2024b), and some finding their creativity to be on par with each other (Góes et al., 2023; Gilhooly, 2024). However, past research has also found that the high LLM performance can be attributed to the artificial nature of the creativity tasks (Ismayilzada et al., 2024a) commonly employed to evaluate LLMs, such as the Alternative Uses Task (Guilford, 1967) or to the remarkable creativity of human-written texts on the web (Lu et al., 2024). Consequently, LLMs have been shown to often lack novelty and surprise in their generations (Ismayilzada et al., 2024a,b; Zhang et al., 2025; Tian et al., 2024; Chakrabarty et al., 2024) and produce significantly less diverse content compared to humans (Padmakumar and He, 2023; Anderson et al., 2024; Kirk et al., 2023; Xu et al., 2024; O'Mahony et al., 2024; Zhang et al., 2024; Wenger and Kenett, 2025). These tendencies limit the utility of LLMs for creative tasks, such as story generation and creative problem solving that often require longer responses and "out-of-the-box" thinking (Tian et al., 2023; Huang et al., 2024; Chen et al., 2024).

Recent research has proposed some methods for improving the creativity of LLMs, often targeting the diversity aspect alone (Wong et al., 2024; Hayati et al., 2023; Chung et al., 2023; Franceschelli and Musolesi, 2024; Zhang et al., 2024; Wang et al., 2024b; Zhou et al., 2025; Lanchantin et al., 2025; Chung et al., 2025) or focusing on a single creativity task (Tian et al., 2023; Nair et al., 2024; Summers-Stay et al., 2023). However, creativity is a multifaceted ability that also encompasses nov-

[*] Work done while at Pennsylvania State University

[†] Equal supervision

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9580–9609

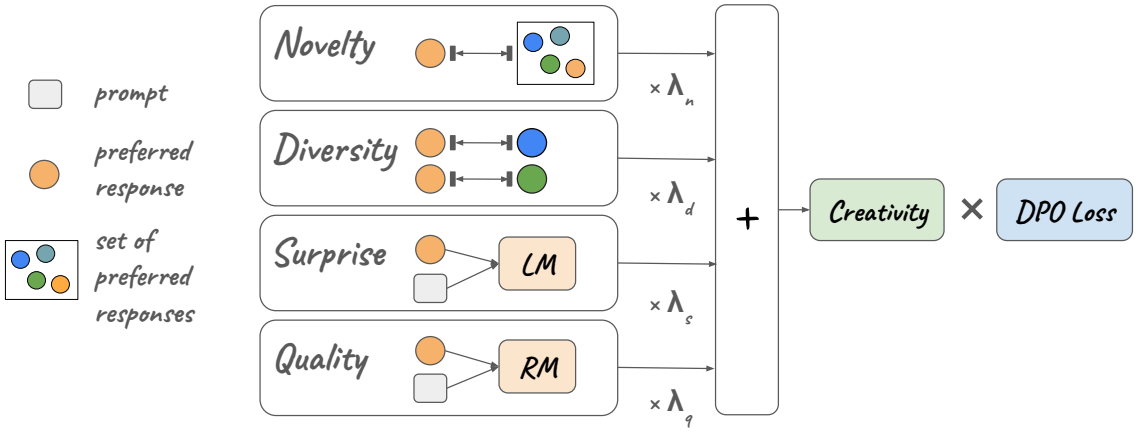

Figure 1: Our preference alignment method CRPO to improve output creativity by injecting a weighted combination of signals from multiple creativity dimensions.

elty, surprise, and quality and manifests itself in a wide range of tasks. Consequently, it has been argued that methods promoting creativity improvements should consider multiple dimensions of creativity together across several creative tasks (Ismayilzada et al., 2024a). Hence, the broader challenge of enhancing overall creativity in LLM outputs largely remains underexplored.

To this end, we propose a novel approach to directly optimize for creativity in language model generation through preference learning (Ouyang et al., 2022; Rafailov et al., 2023). Recent works targeting improvement in LLM creativity have mainly focused on black-box techniques to elicit creative outputs through input-level (e.g., prompting) (Tian et al., 2023; Mehrotra et al., 2024; Nair et al., 2024; Summers-Stay et al., 2023) and output-level strategies (e.g., creative decoding) (Nguyen et al., 2024; Franceschelli and Musolesi, 2024; Meister et al., 2023). However, these methods are inherently limited to the fixed creative capacity of language models and are not designed to optimize for fine-grained dimensions of creativity. Recently, motivated by the negative impact of the preference alignment techniques on the diversity of LLM outputs (Padmakumar and He, 2023; Anderson et al., 2024; Kirk et al., 2023; O'Mahony et al., 2024; West and Potts, 2025), few works have suggested directly modifying the preference optimization methods to promote output diversity (Lanchantin et al., 2025; Chung et al., 2025). Inspired by these approaches, we design a new optimization strategy that *injects* signals from multiple dimensions of creativity into the preference modeling objective in a *modular* fashion. Specifically, we integrate the novelty, diversity, surprise, and quality dimensions of creativity into the training objective of direct preference optimization (DPO) (Rafailov et al., 2023), with weighted composition that allows balancing each dimension's contribution. We call this method *creative preference optimization* (CRPO) and provide its conceptual illustration in Figure 1 with full details in Section 3.

We test the efficacy of CRPO using MUCE (**Mu**ltitask **C**reativity **E**valuation), our newly curated large-scale dataset of prompt-response pairs annotated with human preferences across a diverse range of creative tasks in multiple languages. While previous work has largely evaluated creativity improvements on a narrow range of tasks like story generation (Ismayilzada et al., 2024b; Chung et al., 2025; Lanchantin et al., 2025) or creative problem solving (Tian et al., 2023), MUCE enables us to test whether our methods truly generalize across a diverse range of creativity assessments. Our results show that `Llama-3.1-8B-Instruct` (AI@Meta, 2024) and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023) trained using CRPO outperform the same models trained using only supervised fine-tuning (SFT) or DPO without any creativity injections, as well as existing LLMs such as GPT-4o, generating more novel, diverse, and surprising outputs than all the baselines while maintaining high quality. We publicly release our code, models, and data for future research[1].

Our main contributions are as follows:

1. We introduce MUCE, a large-scale prefer-

[1]https://www.mete.is/creative-preference-optimization/

ence dataset consisting of more than 200,000 human responses and ratings for more than 30 creativity assessments. All tasks within MuCE are carefully chosen to provide valid measures of creativity in humans, making MuCE one of the largest psychologically valid datasets of human creativity for training preference models.

2. We propose a novel flexible preference alignment method CrPO that injects signals from several dimensions of creativity into the existing preference optimization method DPO and train creativity-enhanced versions of `Llama-3.1-8B-Instruct` and `Mistral-7B-Instruct-v0.3`.

3. We evaluate the effectiveness of our approach on a range of creativity tasks from MuCE, as well as external tasks from NoveltyBench (Zhang et al., 2025), using both automated metrics and human evaluations. Our analysis shows that CrPO is a promising method for enhancing the creative capabilities of language models while maintaining quality.

2 Related Work

2.1 Large Language Model Creativity

The potential of building LLM applications for creative industries has spurred significant research interest on AI creativity (Bellemare-Pepin et al., 2024), and many LLM tools marketed for assistance with creative tasks have been developed in the last few years (Wang et al., 2024b). Yet debates on whether AI is capable of true creativity are nearly as old as AI itself (Stein, 2014; Franceschelli and Musolesi, 2024; Sæbø and Brovold, 2024), with theoretical and philosophical arguments being made both for and against AI creativity (Ismayilzada et al., 2024a). Classic psychological theories of creativity generally agree that, for a product to be creative, it must be new, surprising, and valuable (Boden, 2004). Creative tasks are also often characterized by high diversity (Padmakumar and He, 2023; Shypula et al., 2025), though diversity is only one facet of creativity (Johnson et al., 2021). Studies on LLM creativity have yielded conflicting findings: some suggest LLMs surpass human creativity (Bellemare-Pepin et al., 2024; Zhao et al., 2024), others argue they fall short (Koivisto and Grassini, 2023; Chakrabarty et al., 2024; Ismayilzada et al., 2024b), while some conclude that LLM and human creativity are roughly equivalent (Gilhooly, 2024; Stevenson et al., 2022; Góes et al., 2023). Some works have suggested that LLMs lack novelty and surprise in their generations (Ismayilzada et al., 2024a,b; Zhang et al., 2025; Tian et al., 2024; Chakrabarty et al., 2024) and their seemingly remarkable creative outputs may be in large part attributable to the remarkable creativity of human-written texts on the web (Lu et al., 2024). Some recent works have suggested improving the creativity of LLMs through prompting techniques (Tian et al., 2023; Mehrotra et al., 2024; Nair et al., 2024; Summers-Stay et al., 2023) and decoding strategies (Franceschelli and Musolesi, 2024; Meister et al., 2023). In this work, we instead explore directly optimizing language models for creativity using human preferences extracted from responses to creativity assessments.

2.2 Preference Learning

Aligning LLMs to human preferences has proven effective in developing models that are helpful and useful to users, leading to the emergence of numerous preference learning methods (Gao et al., 2024; Ouyang et al., 2022; Rafailov et al., 2023). However, prior work has highlighted a lack of diversity in LLM outputs (Anderson et al., 2024; Lanchantin et al., 2025; Wenger and Kenett, 2025; Padmakumar and He, 2023), with alignment often cited as a contributing factor (West and Potts, 2025). In response, recent research has explored modifications to existing preference modeling techniques aimed at mitigating this reduction in diversity. One notable approach, Diverse Preference Optimization, proposes enhancing preference data creation by selecting preference pairs based on a diversity metric (Lanchantin et al., 2025). Another recent method introduces a modification to the optimization objective itself to incorporate a diversity signal (Chung et al., 2025). Both strategies have demonstrated effectiveness in promoting output diversity with minimal impact on output quality. However, as previously noted, diversity represents only one facet of creativity; true creativity also requires the capacity for novelty and surprise. In this work, we present a modular preference alignment framework for creativity that enables direct optimization across multiple dimensions of creative expression.

3 Creative Preference Optimization

According to its three-criterion definition, creativity involves the generation of novel, high-quality, and

surprising ideas (Simonton, 2012; Boden, 2004; Runco and Jaeger, 2012). Moreover, creative outputs tend to be highly diverse across individuals (Anderson et al., 2024). Therefore, to promote overall creativity in LLM outputs, we propose to inject unsupervised metrics related to each dimension of creativity into the loss functions of standard preference optimization methods. We use direct preference optimization (DPO) (Rafailov et al., 2023) to illustrate our modifications to the loss function. Recall that in the standard formulation of DPO, a policy model (p_θ) is directly optimized on a dataset of (x, y^w, y^l) where x, y^w and y^l refer to the model input (i.e. prompt), preferred (i.e. chosen) model response and dispreferred (i.e. rejected) model response, respectively. Using the ratio between the policy model's likelihood and that of the reference SFT model (p_{SFT}) as an implicit reward, the training objective of DPO is defined as follows (Rafailov et al., 2023):

$$l_{DPO} = \left[\log \sigma\left(\beta \log \frac{p_\theta(y^w|x)}{p_{\text{SFT}}(y^w|x)} - \beta \log \frac{p_\theta(y^l|x)}{p_{\text{SFT}}(y^l|x)}\right)\right]$$
$$\mathcal{L}_{\mathcal{DPO}} = -\mathbb{E}_{(x,y^w,y^l)\in D} \quad l_{DPO} \tag{1}$$

A challenge with standard preference optimization methods is that they may significantly reduce the diversity of the responses LLMs generate, as the loss function encourages models to generate preferred responses even if they are not very creative (West and Potts, 2025; Padmakumar and He, 2023; Anderson et al., 2024; Kirk et al., 2023; Xu et al., 2024; O'Mahony et al., 2024; Zhang et al., 2024; Wenger and Kenett, 2025). Existing approaches to address this in the preference optimization objective have centered around curating preference data based on various diversity metrics (Lanchantin et al., 2025) or incorporating extra regularization terms that encourage diverse generations while balancing quality (Chung et al., 2025). For example, the recently proposed Diversified DPO (DDPO) method adds a scalar diversity term δ^w (i.e. diversity score of the preferred response) into the DPO loss (Chung et al., 2025):

$$\mathcal{L}_{\mathcal{DDPO}} = -\mathbb{E}_{(x,y^w,y^l)\in D} \quad \delta^w l_{DPO} \tag{2}$$

While diversity is important for creativity, research in psychology has long established that truly creative responses also require novelty, surprise, and quality (Boden, 2004; Barron, 1955; Simonton, 2018). Therefore, we propose incorporating metrics for each of these, alongside diversity, into the preference loss in a *modular* structure, enabling the construction of different creativity models by combining these dimensions as needed.

$$\mathcal{L}_{\mathcal{CDPO}} = -\mathbb{E}_{(x,y^w,y^l)\in D}\Big[(\lambda_d\delta^w + \lambda_n\nu^w + \lambda_s\xi^w + \lambda_q\gamma^w) l_{DPO}\Big] \tag{3}$$

In our proposed creative DPO loss, δ^w, ν^w, ξ^w and γ^w correspond to diversity, novelty, surprise and quality scores of the preferred response respectively and λ_d, λ_n, λ_s and λ_q are hyperparameters that control the effect of each score (we call them injection weights). In particular, when $\lambda_d = 1$, $\lambda_n = 0$, $\lambda_s = 0$ and $\lambda_q = 0$, we recover the DDPO loss. While there are multiple approaches for operationalizing δ^w, ν^w, ξ^w and γ^w, we propose to use the following metrics for each:

3.1 Diversity

Diversity is defined as the pairwise differences between artifacts, and semantic distance is typically used to measure the difference (Chung et al., 2025; Ismayilzada et al., 2024b; Padmakumar and He, 2023). We use an inverse homogenization metric from Padmakumar and He (2023) similar to Chung et al. (2025). Specifically, given a prompt x and a set of (preferred) responses for x denoted as Y_x, we compute the diversity score of any particular preferred response as the average pairwise semantic distance to all the other preferred responses in Y_x:

$$\delta^w = \frac{1}{|Y_x| - 1} \sum_{y_i \in Y_x \setminus y^w} \textit{semdis}(y^w, y_i) \tag{4}$$

We use $1 - cos\_sim(\cdot,\cdot)$ as a semantic distance function (i.e., $semdis(\cdot,\cdot)$).

3.2 Novelty

Novelty is typically defined as the measure of how different an artifact is from other known artifacts in its class (Maher, 2010), and semantic distance-based metrics have been established as a good proxy in creativity research (Johnson et al., 2022; Beaty and Johnson, 2021; Dunbar and Forster, 2009; Harbinson and Haarman, 2014; Karampiperis et al., 2014). We use a novelty metric similar to Karampiperis et al. (2014) where the novelty of a text is defined as the absolute difference between the average pairwise semantic distances of words in the text and those of a reference corpus of texts. In particular, we define the set of preferred responses to a prompt x as the reference corpus (Y_x) and define the novelty of a preferred response as follows:

$$\nu^w = |DSI(y^w) - DSI(Y_x)| \quad (5)$$

$$DSI(T) = \frac{\sum_{i,j=1}^{|T|} semdis(T_i, T_j),\ i \neq j}{|T|} \quad (6)$$

Here T refers to a piece of text, T_i to the word i in the set of unique words in T denoted as $|T|$, and $DSI(\cdot)$ is *divergent semantic integration*, the average pairwise semantic distances of words in a text (Johnson et al., 2022).

3.3 Surprise

Surprise or unexpectedness has many definitions in the cognitive science literature (Modirshanechi et al., 2022), but it is generally characterized by deviation from the expected. We use Shannon surprise – the negative log-likelihood of the text — which has been widely used as a measure of surprise in prior work (Bunescu and Uduehi, 2022; Modirshanechi et al., 2022; Kuznetsova et al., 2013). More specifically, given a prompt x, we define the surprise of a particular response as the exponentiated negative log-likelihood of the response (i.e. perplexity) conditioned on the prompt x and under some reference model S as follows:

$$\xi^w = 2^{-logP_S(y^w|x)} \quad (7)$$

3.4 Quality

Although a general quality scoring method is hard to define as it is highly domain-dependent, reward models that are trained to output a high score to preferred answers are increasingly being used to assess the overall quality of language model outputs and align them to human preference (Zhang et al., 2025; Lambert et al., 2024). In particular, we define the quality of a preferred response given a prompt x as the score assigned by some reward model R: $\gamma^w = R(y^w|x)$.

4 The MuCE Dataset

To compile MuCE, we solicited data from the global creativity research community, specifically targeting researchers studying human creativity to obtain data from tasks known to be valid creativity measures. We specifically targeted datasets which contained complete metadata, including information about the task, language, and items that participants responded to. We gathered additional data by performing a manual search of the Open Science Framework database[2], and only retained data from peer-reviewed articles. In total, 43% of the data in MuCE has never been publicly released, making it unlikely that LLMs have seen the item-response combinations for the majority of our tasks.

Every response in MuCE was rated for creativity by at least two raters, and in some cases up to 75 employing a missing-raters design (Forthmann et al., 2025). While it is common practice to measure creativity using multiple independent raters, individual raters may deliver unhelpful or noisy ratings if they did not understand the task instructions, had a different understanding of the rating criteria, or for other reasons (Forthmann et al., 2017). To account for this, we followed best practices for subjective scoring tasks by employing Judge Response Theory (Myszkowski and Storme, 2019) to check for raters whose ratings were uninformative in an information-theoretic sense. We fit JRT models to each task within MuCE, which gave us an information function for each rater across tasks. We then input the results from the JRT into a genetic algorithm (Schroeders et al., 2016) which identified a subset of raters per dataset that maximized the per dataset rater information function.[3] This process dropped uninformative raters from each dataset, enhancing the quality of the final creativity ratings. The individual rater's scores were aggregated via factor scores, as is best practice in creativity assessment (Silvia, 2011), and we rescaled the factor-transformed creativity scores into the integer range 10-50 as is done for prior work in automated creativity assessment (Organisciak et al., 2023). Full details about the dataset construction are in Appendix A.

5 Experiments

5.1 SFT and Preference Datasets

While our MuCE dataset contains samples for multiple languages, we focus on showing the effectiveness of CrPO on the English subset in this work and leave experiments using the full dataset as future work. From the base English MuCE dataset, we generate a preference dataset by creating tuples of preferred and rejected responses to the same prompt, treating the response that received the higher creativity score as the preferred one. Past work has shown that data quality is one of the main factors behind preference model performance (Liu et al., 2024; Deng et al., 2025; Wang et al., 2024a).

[2] https://osf.io/

[3] While ensuring that the algorithm kept at least two raters per dataset.

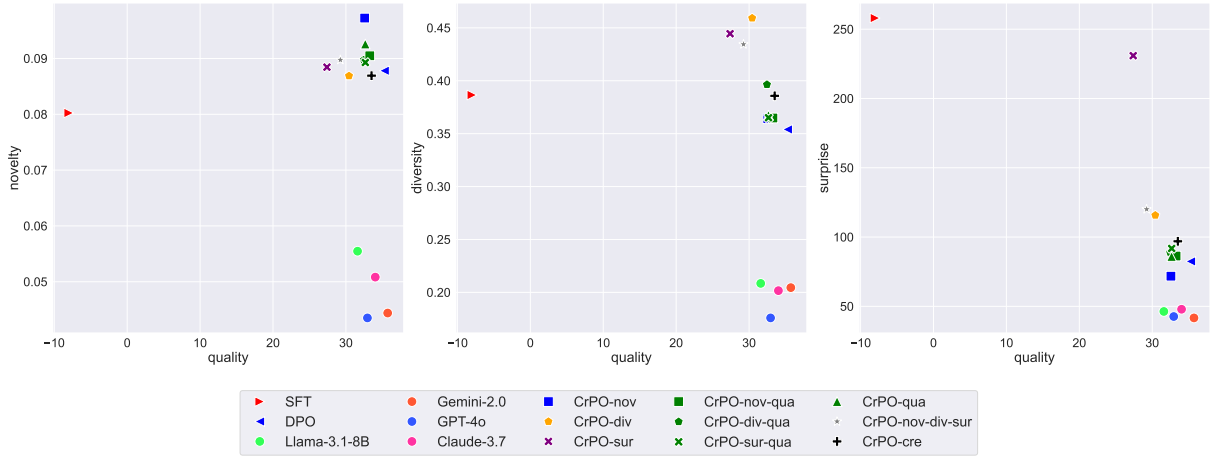

Figure 2: Results on held-out evaluation suite from MuCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3` results can be found in Appendix Figure 6.

Therefore, we curate a high-quality SFT dataset of $5,285$ samples (MuCE-SFT) and preference dataset of $42,058$ samples (MuCE-Pref) from the base MuCE which we detail in Appendix B.

5.2 Training

Models As our base models, we use `Llama-3.1-8B-Instruct` (AI@Meta, 2024) and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023) and implement CrPO as described in Section 3. We first train our models using supervised fine-tuning (SFT model) for a single epoch on MuCE-SFT, and then apply preference optimization on the SFT model using CrPO and MuCE-Pref dataset. We train all models using parameter-efficient tuning with LoRA using a rank of 128 and an alpha of 256 (Hu et al., 2022). Additional details on the training setup can be found in Appendix C.

Creativity Injection We compute creativity metric scores for each preferred response and inject them into the DPO objective function as described in Section 3. Since each metric is on a different scale and we would like to combine the effects of different injections, we normalize each score to a range of $[0, 1]$ before injection. We vary the injection weights $\lambda_d, \lambda_n, \lambda_s, \lambda_q$ accordingly[4] to train different suites of creative models. As novelty and diversity measures require a reference set to compute against, we adopt a prompt-level granularity and consider the set of responses for a given prompt as the reference corpus similar to prior work (Chung et al., 2025). We use the `jina-embeddings-v3` model (Sturua et al., 2024) to compute text embeddings for all metrics that rely on semantic distance. For surprise, we use instruction-tuned `Gemma-2-27B` (Google, 2024a) as our reference surprise model S. While our creativity preference dataset is already high-quality, we also experiment with injecting external quality signals to study its interaction with other creativity dimensions. Hence, for the quality measure, we employ an existing reward model `Skywork-Reward-Gemma-27B-v0.2` (Liu et al., 2024) that is one of the top-performing models on RewardBench (Lambert et al., 2024) as our reference reward model R.

5.3 Evaluation

Tasks and Metrics We evaluate all models across several dimensions of creativity on held-out prompts of various tasks and two held-out tasks. More specifically, we use 5 held-out prompts from *Real-Life Creative Problem Solving, Alternate Uses of Objects, Design Solutions, Hypothesis Generation,* and *Metaphors* tasks, and 9 prompts from two held-out tasks of *Poems* and *Sentence Completion*. For each prompt, we generate 16 responses from each model by varying the $temperature$, top_p, and top_k decoding parameters. Our final held-out evaluation suite contains 224 samples. We evaluate the responses on the dimensions of novelty, diversity, and surprise using the metrics described in Sec-

[4]For example, to train a novelty model, we set $\lambda_n = 1$ and others to 0 whereas for novelty and quality model we set $\lambda_n = 1$ and $\lambda_q = 1$.

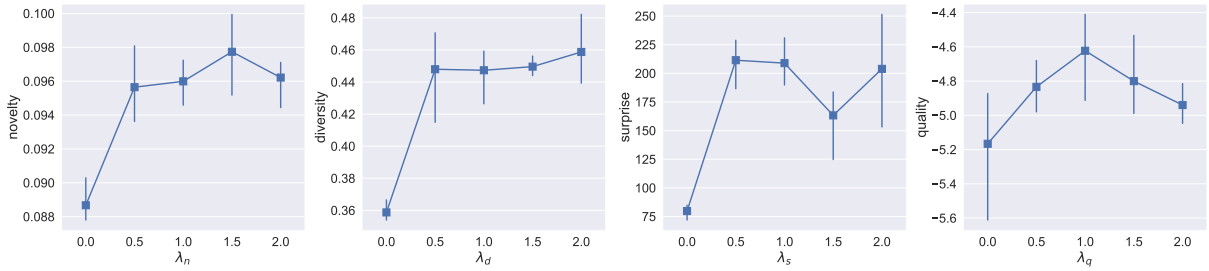

Figure 3: Effect of injection weights for each dimension. Results are averaged across three seed runs.

tion 3. Additionally, to study the tradeoff between creativity and quality, we train a reward model on our preference dataset using instruction tuned `Gemma-2-9b` (Google, 2024a) and use it to score the overall quality of model generations. More details about the evaluation setup can be found in Appendix D.

Baselines As baselines [5], we use the base models `Llama-3.1-8B-Instruct` (AI@Meta, 2024) and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), SFT models which are the base models supervised fine-tuned on MUCE-SFT, a vanilla DPO model trained on top of the SFT model using the MUCE-PREF dataset without any creativity injections and three closed-source instruction-tuned LLMs, namely `GPT-4o` (OpenAI, 2024), `Claude-3.7-Sonnet` (Anthropic, 2025), and `Gemini-2.0-Flash` (Google, 2024b). We also consider additional baselines such as *"Brainstorm, then select"*, a creative prompting approach that has shown improvement on creativity scores in the Alternative Uses of Objects Task (AUT) (Summers-Stay et al., 2023) and *min-p sampling*, a decoding strategy that directly promotes output creativity (Nguyen et al., 2024). Evaluation details for these baselines can be found in Appendix D.3.

CRPO Models We train several CRPO models corresponding to the different dimensions of creativity. More specifically, for each dimension, we train a model that is injected with a signal for the given dimension and another model that is injected with a signal for both the given dimension (e.g. CRPO-`nov`) and the quality dimension (e.g. CRPO-`nov-qua`). We train the latter models to understand the tradeoff between other dimensions of creativity and the quality that has been reported in previous research (Zhang et al., 2025; Lanchantin et al., 2025; Chung et al., 2025). Additionally, we train two creative models that inject all dimensions of creativity (denoted as CRPO-`cre`) and all except quality (denoted as CRPO-`nov-div-sur`). In all these experiments, λ injection weights are set to 1 for simplicity. We perform a more detailed analysis of these hyperparameters in Section 6.1.

6 Results

Figure 2 summarizes performance on our held-out evaluation suite across creativity dimensions for all baselines and CRPO models using the `Llama-3.1-8B-Instruct` as a base. Results for `Mistral-7B-Instruct-v0.3` can be found in Appendix Figure 6 and follows the same trends. First, we observe a clear separation between existing instruction-tuned LLMs and our models: while the former cluster around high quality but low novelty, diversity, and surprise, our models achieve high scores across all four dimensions. Second, for each creativity dimension, the model trained with that specific injection outperforms others on the same metric, confirming the effectiveness of targeted optimization, without a considerable drop in quality.

Models that combine a creativity signal with an external quality signal (CRPO-`{nov,div,sur}-qua`) improve in quality but show reduced performance on the targeted dimension, illustrating a trade-off. The same pattern holds when comparing the CRPO-`nov-div-sur` model to the full CRPO-`cre` model, further highlighting the balance between quality and other facets of creativity. Interestingly, the vanilla DPO model, without any creativity injections, already outperforms existing LLM baselines, demonstrating the strength of our preference dataset. Still, most of our creativity-optimized models

[5] We note that DDPO (Chung et al., 2025) is a special case of our method, and the model trained with it is equivalent to our CRPO-`div` variant. For this reason, we do not treat it as a separate baseline. Moreover, since DDPO has already been shown to outperform the Diverse Preference Optimization method (Lanchantin et al., 2025), a direct comparison is unnecessary by transitivity.

significantly surpass DPO across all dimensions. Finally, the SFT model performs worst in quality and shows only comparable performance on other dimensions, reinforcing prior findings (Chung et al., 2025) about the limited generalizability of supervised fine-tuning in creative tasks, where no single "correct" answer exists.

We also compare our approach to creative prompting (Brainstorm & Select) and decoding (*min-p* sampling) strategies which we detail in Appendix D.3 and results are reported in Appendix Tables 2 and 6. We can see that while the creative prompting and decoding strategies improve baseline performance, our models with the standard prompting and decoding approaches still beat them across all dimensions with minimal drop in quality.

Overall, **our results show that CrPO enhances multiple aspects of creativity with minimal impact on quality**, offering a flexible and effective framework for creativity alignment in LLMs.

6.1 Effect of Injection Weights

While we set all injection weights to 1 for simplicity in our main evaluations, we also study the effect of the different injection values on the performance of models across dimensions. In particular, we vary the injection weights from 0 to 2.0 with an increment of 0.5 for all dimensions and report the averaged results across three seed runs in Figure 3. We observe that across most dimensions, an injection weight of 0.5 yields the greatest performance gains, with further increases resulting in diminishing returns or slight performance degradation. In terms of quality, the injection weight of 1.0 results in the highest performance. To gain further insights into the inherent interactions between creativity dimensions, for each dimension, we also report the performance of a model optimized for a given dimension on all the other dimensions and across increasing values of injection weights. Results for this analysis can be found in Appendix Figures 8, 9, 10. Results show that in general, the interaction between different creativity dimensions is complex with some discernible patterns. One clear trend is that the increase in novelty, diversity, and surprise injections results in a quality drop, though not significant. Interestingly, we observe that while an increase in novelty injection follows with overall increase in surprise performance, the opposite does not seem to be true and in fact, more surprise injection seems to cause a drop in novelty. On the other hand, novelty and diversity dimensions seem to have a mutual positive correlation. Finally, we observe no clear relationship between the dimensions of surprise and diversity. In general, we suggest tuning the injection weights depending on the training dataset, underlying task, and the base model for the best performance.

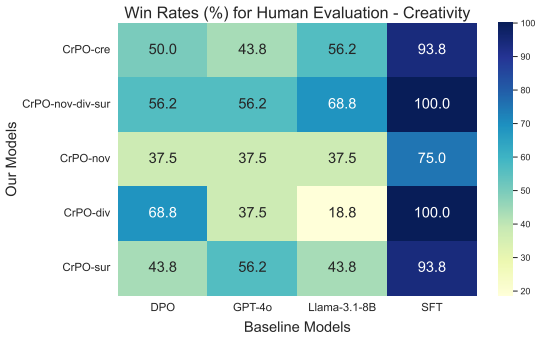

Figure 4: Human evaluation results measured by win rates. Participants were asked to make a pairwise comparison between our models and baselines with respect to the overall creativity.

6.2 Human Evaluation

In addition to automated metrics, we conduct a human evaluation to assess the real-world effectiveness of our approach. Due to the high cost of human studies, we focus on the overall creativity dimension using a single task (Sentence Completion), 4 prompts, 4 baselines (SFT, DPO, `Llama-3.1-8B-Instruct`, and `GPT-4o`), and 5 CRPO variants (`nov`, `div`, `sur`, `nov-div-sur`, `cre`). In a blind pairwise setup, participants compared responses from a baseline and a CRPO model for creativity, unaware that the texts were AI-generated. A total of 320 comparisons were collected with balanced sampling across models. Additional details are in Appendix D.1.

Figure 4 presents the win rates. The CRPO-`nov-div-sur` model consistently outperforms all baselines, particularly `Llama-3.1-8B-Instruct`, by a wide margin. In contrast, the full CRPO-`cre` model lags slightly, reflecting the creativity–quality tradeoff seen in automated evaluations. Notably, CRPO models achieve especially strong gains over SFT, reinforcing previous findings.

6.3 NOVELTYBENCH Evaluation

While we demonstrate the effectiveness of our approach on the MUCE held-out set using automated metrics, we also evaluate generalization on external benchmarks using the recently introduced NOVEL-

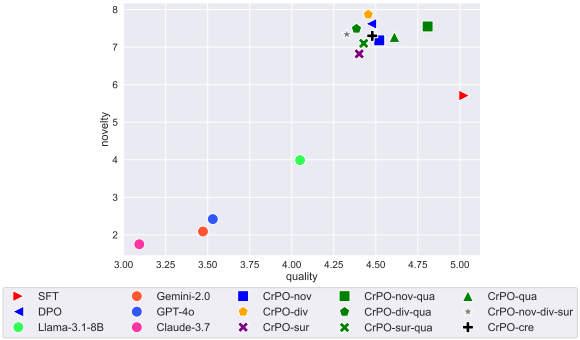

Figure 5: Evaluation results on NOVELTYBENCH, using the novelty and quality metrics defined in Zhang et al. (2025).

TYBENCH (Zhang et al., 2025). This benchmark includes tasks spanning randomness, factual knowledge, creative writing, and subjectivity. Following the recommended evaluation setup, we benchmark all baselines and CRPO variants on a curated 100-prompt subset, using the benchmark's novelty and quality metrics. Full details are in Appendix D.2.

Figure 5 shows novelty vs. quality scores across all models and tasks. As in our internal evaluation, we observe a clear separation: existing LLM baselines cluster around lower novelty and variable quality, while our models consistently achieve high scores on both dimensions. Notably, although our models outperform SFT on novelty, the SFT model surprisingly achieves higher quality—beating both baselines by a large margin and our models by a smaller one. This aligns with findings from NOVELTYBENCH (Zhang et al., 2025), where smaller models like `Gemma-2-2B-it` and `Llama-3.1-8B-Instruct` often surpass larger ones in quality.

Overall, our models set a new state-of-the-art on the NOVELTYBENCH leaderboard in terms of novelty. [6]

7 Discussion

In this work, we propose a modular preference alignment framework that uses signals from different dimensions of creativity to promote overall output creativity improvement. We show that small LLMs trained using our framework on a psychologically-valid datasets of human creativity outperform strong baselines such as GPT-4o on both automated and human evaluations. Our formulation builds on well-established dimensions from creativity research—specifically novelty, surprise, diversity, and quality—which are frequently identified as core components in both cognitive psychology (Modirshanechi et al., 2022; Grace and Maher, 2016; Barron, 1955) and computational creativity literature (Runco and Jaeger, 2012; Simonton, 2012; Boden, 2004). We also employ creativity metrics that provide measurable signals aligning with key cognitive theories and enable practical optimization within language models. These signals have also been shown to provide a good feedback mechanism and increase human creativity in several tasks (De Chantal and Organisciak, 2023; de Chantal et al., 2025b). Similarly, our work shows that these signals can be used as an implicit feedback mechanism in the learning process to improve language model creativity. Additionally, our framework is well-suited for downstream creative tasks where certain dimensions of creativity matter more than others. In creative writing, novelty and diversity ensure fresh, engaging content, while surprise drives plot twists. In advertising, novelty and surprise capture audience attention in crowded markets. For brainstorming, diversity is key to exploring a wide idea space. By enabling models to be trained with customizable mixes of creativity dimensions, our framework adapts to the needs of each domain.

8 Conclusion

We introduce CRPO, a flexible methodology for enhancing the creativity of LLMs. Leveraging a novel large-scale human preference dataset focused on creativity, we show that models aligned with CRPO produce generations that are not only novel, diverse, and surprising, but also high in quality —— on both our held-out evaluation suite and the external NOVELTYBENCH dataset. Human evaluations further confirm that raters consistently judge our model's outputs to be more creative than those of several strong baselines, highlighting the potential of our approach to boost LLM creativity. While our experiments focus on smaller models such as `Llama-3.1-8B` and an English-only dataset, future work could explore the scalability of CRPO to larger models, multilingual settings and other preference optimization methods.

Limitations

Due to constraints on both computational resources and budget for human studies, we were unable to evaluate CRPO on any languages other than En-

[6] https://novelty-bench.github.io/

glish. Multilingual creativity assessment using generative AI remains a challenging problem and an active area of research (Haase et al., 2025). While we believe our data represents a valuable resource for the community, future work will need to test our methods in multilingual settings to ensure multilingual generalization. These compute constraints also prevented us from evaluating CRPO on larger open-weight models, making scaling trends difficult to predict. We retained only samples with full agreement for the creativity score when training our models. While this aligns with best practices for creativity measurement in psychology (Cseh and Jeffries, 2019), it may also mask genuine sources of rater disagreement that should be modeled. Moreover, we acknowledge that, much like other datasets used to align LLMs, the preferences represented by our annotator population likely do not reflect the full range of human preferences, which could bias our models' generations (Yeh et al., 2024). We believe that the large-scale and multilingual nature of our collected data likely makes it one of the most representative creativity datasets currently available, but stress that future work should consider issues of bias and fairness more carefully for LLM creativity assessment. Finally, we also acknowledge that creativity is a complex and subjective construct, and the metrics we used may not capture the full richness of human creativity judgments.

Ethical Considerations

We emphasize that our models should not be used for safety-critical applications, as the relationship between creativity and alignment with other values remains underexplored. Notably, our dataset contains responses to tests of malevolent creativity that are by definition unsafe for models to generate. We also observed qualitatively that CRPO models were more likely to generate unsafe or toxic responses even to prompts that did not explicitly request such behaviors. We believe that our data is valuable for red-teaming evaluations on tasks requiring creativity, and that aligning models on these malevolent responses could be beneficial for understanding how malicious actors might use creativity-enhanced models to execute unsafe goals. However, we also acknowledge the ethical concerns that the release of our models and datasets would raise, and believe that restricting access to only those which have signed a license agreement is the best approach for balancing safety with continued scientific advancement. While we believe our results demonstrate how aligning LLMs with carefully designed human creativity datasets can significantly improve the novelty and diversity of their generations, it remains unclear how to both optimize for creativity while preserving guardrails that prevent unsafe behavior.

We also acknowledge the broader debates around the valid use of AI in social-behavioral research (Sun et al., 2025) and concerns surrounding AI automation of industries requiring creativity (Wilkinson, 2023) in which our work is situated. While the over-reliance on AI for creative tasks to the detriment of human welfare is a legitimate concern, AI has also been acknowledged for its potential to enhance human creativity above and beyond what might be possible otherwise (de Chantal et al., 2025a; Loi and van der Plas, 2020; Loi et al., 2020). Creativity is a vital skill for future knowledge workers to master (Forum, 2025), and we believe that enhancing the creativity of AI is an important prerequisite for developing AI systems capable of training humans to be more creative.

Acknowledgements

Mete and Lonneke gratefully acknowledge the support of the Swiss National Science Foundation (grant 205121_207437: C - LING) and Fondazione Aldo e Cele Daccò. Antoine Bosselut gratefully acknowledges the support of the Swiss National Science Foundation (No. 215390), Innosuisse (PFFS-21-29), the EPFL Center for Imaging, Sony Group Corporation, and a Meta LLM Evaluation Research Grant. Roger E. Beaty is supported by grants from the US National Science Foundation [DRL-1920653; DRL-240078; DUE-2155070].

References

Sergio Agnoli, Giovanni E Corazza, and Mark A Runco. 2016. Estimating creativity with a multiple-measurement approach within scientific and artistic domains. *Creativity Research Journal*, 28(2):171–176.

AI@Meta. 2024. Llama 3 model card.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Anthropic. 2025. Claude 3.7 sonnet and claude code.

Frank Barron. 1955. The disposition toward originality. *The Journal of Abnormal and Social Psychology*, 51(3):478.

Roger Beaty, Robert A Cortes, Simone Luchini, John D Patterson, Boris Forthmann, Brendan S Baker, Baptiste Barbot, Mariale Hardiman, and Adam Green. 2024. The scientific creative thinking test (sctt): Reliability, validity, and automated scoring. *PsyArxiv Preprints*.

Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior research methods*, 53(2):757–780.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*.

Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Razvan C. Bunescu and Oseremen O. Uduehi. 2022. Distribution-based measures of surprise for creative language: Experiments with humor and metaphor. *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Soma Chaudhuri, Alan Pickering, and Joydeep Bhattacharya. 2025. Evaluating poetry: Navigating the divide between aesthetical and creativity judgments. *The Journal of Creative Behavior*, 59(1):e683.

Qi Chen, Bowen Zhang, Gang Wang, and Qi Wu. 2024. Weak-eval-strong: Evaluating and eliciting lateral thinking of llms with situation puzzles. *arXiv preprint arXiv:2410.06733*.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. 2025. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*.

Katherine N Cotter, Jean E Pretz, and James C Kaufman. 2016. Applicant extracurricular involvement predicts creativity better than traditional admissions factors. *Psychology of Aesthetics, Creativity, and the Arts*, 10(1):2.

Genevieve M Cseh and Karl K Jeffries. 2019. A scattered cat: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):159.

Pier Luc de Chantal, Roger Beaty, Antonio Laverghetta, Jimmy Pronchick, John Patterson, Peter Organisciak, Katarzyna Potega vel Zabik, Baptiste Barbot, and Maciej Karwowski. 2025a. Artificial intelligence enhances human creativity through real-time evaluative feedback.

Pier Luc de Chantal, Roger Beaty, Antonio Laverghetta, Jimmy Pronchick, John Patterson, Peter Organisciak, Katarzyna Potega vel Zabik, Baptiste Barbot, and Maciej Karwowski. 2025b. Artificial intelligence enhances human creativity through real-time evaluative feedback.

Pier-Luc De Chantal and Peter Organisciak. 2023. Automated feedback and creativity: On the role of metacognitive monitoring in divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2025. Less is more: Improving llm alignment via preference data selection. *arXiv preprint arXiv:2502.14560*.

Paul V DiStefano, John D Patterson, and Roger E Beaty. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, pages 1–15.

Paul V DiStefano, Daniel Zeitlen, Janet Rafner, Pier-Luc de Chantal, Aoran Peng, Scarlett Miller, and Roger Beaty. 2025. Evaluating ai's ideas: The role of individual creativity and expertise in human-ai co-creativity.

Kevin Dunbar and Eve Forster. 2009. Creativity evaluation through latent semantic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Li Fan, Kaixiang Zhuang, Xueyang Wang, Jingyi Zhang, Cheng Liu, Jing Gu, and Jiang Qiu. 2023. Exploring the behavioral and neural correlates of semantic distance in creative writing. *Psychophysiology*, 60(5):e14239.

Boris Forthmann, Benjamin Goecke, and Roger E Beaty. 2025. Planning missing data designs for human ratings in creativity research: A practical guide. *Creativity Research Journal*, 37(1):167–178.

Boris Forthmann, Heinz Holling, Nima Zandi, Anne Gerwig, Pınar Çelik, Martin Storme, and Todd Lubart. 2017. Missing creativity: The effect of cognitive workload on rater (dis-) agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, 23:129–139.

World Economic Forum. 2025. Future of jobs report.

Giorgio Franceschelli and Mirco Musolesi. 2024. Creative beam search: Llm-as-a-judge for improving response generation. *arXiv preprint arXiv:2405.00099*.

Bofei Gao, Feifan Song, Yibo Miao, Zefan Cai, Zhe Yang, Liang Chen, Helan Hu, Runxin Xu, Qingxiu Dong, Ce Zheng, Wen Xiao, Ge Zhang, Daoguang Zan, Keming Lu, Bowen Yu, Dayiheng Liu, Zeyu Cui, Jian Yang, Lei Sha, and 5 others. 2024. Towards a unified view of preference learning for large language models: A survey. *Preprint*, arXiv:2409.02795.

Ken Gilhooly. 2024. Ai vs humans in the aut: Simulations to llms. *Journal of Creativity*, 34(1):100071.

Benjamin Goecke, Paul V DiStefano, Wolfgang Aschauer, Kurt Haim, Roger Beaty, and Boris Forthmann. 2024a. Automated scoring of scientific creativity in german. *The Journal of Creative Behavior*, 58(3):321–327.

Benjamin Goecke, Selina Weiss, and Oliver Wilhelm. 2024b. Driving factors of individual differences in broad retrieval ability: Gr is more than the sum of its parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Fabrício Góes, Piotr Sawicki, Marek Grzes, Marco Volpe, and Jacob Watson. 2023. Pushing gpt's creativity to its limits: Alternative uses and torrance tests. In *ICCC*.

Google. 2024a. Gemma 2: Improving open language models at a practical size.

Google. 2024b. Introducing gemini 2.0: our new ai model for the agentic era.

Kazjon Grace and Mary Lou Maher. 2016. Surprise-triggered reformulation of design goals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

J.P. Guilford. 1967. *The Nature of Human Intelligence*. McGraw-Hill series in psychology. McGraw-Hill.

Jennifer Haase, Paul H. P. Hanel, and Sebastian Pokutta. 2025. S-dat: A multilingual, genai-driven framework for automated divergent thinking assessment. *Preprint*, arXiv:2505.09068.

J Harbinson and Henk Haarman. 2014. Automated scoring of originality using semantic representations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? *arXiv preprint arXiv:2311.09799*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Ruizhi He, Kaixiang Zhuang, Lijun Liu, Ke Ding, Xi Wang, Lei Fu, Jiang Qiu, and Qunlin Chen. 2022. The impact of knowledge on poetry composition: An fmri investigation. *Brain and language*, 235:105202.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. 2024. LatEval: An interactive LLMs evaluation benchmark with incomplete information from lateral thinking puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10186–10197, Torino, Italia. ELRA and ICCL.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024a. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024b. Evaluating creative short story generation in humans and large language models. In *Proceedings of ICCC 2025*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Dan R Johnson, Andrew S Cuthbert, and Mara E Tynan. 2021. The neglect of idea diversity in creative idea generation and evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, 15(1):125.

Dan Richard Johnson, J. Kaufman, Brendan S. Baker, John D. Patterson, Baptiste Barbot, Adam E. Green, Janet G. van Hell, Evan S. Kennedy, Grace F Sullivan, Christa L. Taylor, Thomas Ward, and Roger E. Beaty. 2022. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 55:3726 – 3759.

Hansika Kapoor, Hreem Mahadeshwar, Sarah Rezaei, Roni Reiter-Palmon, and James C Kaufman. 2024. The ties that bind: Low morals, high deception, and dark creativity. *Creativity Research Journal*, pages 1–20.

Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1):13601.

Polina Kuznetsova, Jianfu Chen, and Yejin Choi. 2013. Understanding and quantifying creativity in lexical composition. In *Conference on Empirical Methods in Natural Language Processing*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Michele Loi and Lonneke van der Plas. 2020. A blindspot of ai ethics: anti-fragility in statistical prediction. In *Proceedings of the SwissText 2020*.

Michele Loi, Eleonora Viganò, and Lonneke van der Plas. 2020. The societal and ethical relevance of computational creativity. In *Proceedings of the ICCC 2020*.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and 1 others. 2024. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *arXiv preprint arXiv:2410.04265*.

Simone A Luchini, Nadine T Maliakkal, Paul V DiStefano, Antonio Laverghetta Jr, John D Patterson, Roger E Beaty, and Roni Reiter-Palmon. 2025. Automated scoring of creative problem solving with large language models: A comparison of originality and quality ratings. *Psychology of Aesthetics, Creativity, and the Arts*.

Mary Lou Maher. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, pages 22–28.

Pronita Mehrotra, Aishni Parab, and Sumit Gulwani. 2024. Enhancing creativity in large language models through associative thinking strategies. *arXiv preprint arXiv:2405.06715*.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Alireza Modirshanechi, Johanni Brea, and Wulfram Gerstner. 2022. A taxonomy of surprise definitions. *Journal of Mathematical Psychology*, 110:102712.

Nils Myszkowski and Martin Storme. 2019. Judge response theory? a call to upgrade our psychometrical account of creativity judgments. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):167.

Lakshmi Nair, Evana Gizzi, and Jivko Sinapov. 2024. Creative problem solving in large language and vision models - what would it take? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11978–11994, Miami, Florida, USA. Association for Computational Linguistics.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2024. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

John D Patterson, Hannah M Merseal, Dan R Johnson, Sergio Agnoli, Matthijs Baas, Brendan S Baker, Baptiste Barbot, Mathias Benedek, Khatereh Borhani, Qunlin Chen, and 1 others. 2023. Multilingual semantic distance: Automatic verbal creativity assessment in many languages. *Psychology of Aesthetics, Creativity, and the Arts*, 17(4):495.

Corinna Perchtold-Stefan, Hansika Kapoor, James C Kaufman, Hreem Mahadeshwar, and Alison Fernandes. 2024. Development and neuronal validation of the dark creativity deception battery (dcdb).

Corinna M Perchtold-Stefan, Christian Rominger, Ilona Papousek, and Andreas Fink. 2023. Functional eeg alpha activation patterns during malevolent creativity. *Neuroscience*, 522:98–108.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Tuval Raz, Simone Luchini, Roger Beaty, and Yoed Kenett. 2024. Bridging the measurement gap: A large language model method of assessing open-ended question complexity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity research journal*, 24(1):92–96.

Solve Sæbø and Helge Brovold. 2024. On the stochastics of human and artificial creativity. *arXiv preprint arXiv:2403.06996*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *Preprint*, arXiv:2310.10076.

Janika Saretzki, Rosalie Andrae, Boris Forthmann, and Mathias Benedek. 2024. Investigation of response aggregation methods in divergent thinking assessments. *The Journal of Creative Behavior*.

Ulrich Schroeders, Oliver Wilhelm, and Gabriel Olaru. 2016. Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PloS one*, 11(11):e0167110.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025. Evaluating the diversity and quality of llm generated content. *arXiv preprint arXiv:2504.12522*.

Paul J Silvia. 2011. Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity*, 6(1):24–30.

Dean Keith Simonton. 2012. Taking the us patent office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity research journal*, 24(2-3):97–106.

Dean Keith Simonton. 2018. Defining creativity: Don't we also need to define what is not creative? *The Journal of Creative Behavior*, 52(1):80–90.

Morris I Stein. 2014. *Stimulating creativity: Individual procedures*. Academic Press.

Claire E. Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han L. J. van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. In *ICCC*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Douglas Summers-Stay, Stephanie M. Lukin, and Clare R. Voss. 2023. Brainstorm, then select: a generative language model improves its creativity score.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 845–854.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Binghai Wang, Rui Zheng, Lu Chen, Zhiheng Xi, Wei Shen, Yuhao Zhou, Dong Yan, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2024a. Reward modeling requires automatic adjustment based on data quality. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4041–4064.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, and 1 others. 2024b. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Selina Weiss, Benjamin Goecke, and Oliver Wilhelm. 2024. How much retrieval ability is in originality? *The Journal of Creative Behavior*, 58(3):370–387.

Selina Weiss, Sally Olderbak, and Oliver Wilhelm. 2023. Conceptualizing and measuring ability emotional creativity. *Psychology of Aesthetics, Creativity, and the Arts*.

Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*.

Peter West and Christopher Potts. 2025. Base models beat aligned models at randomness and creativity. *Preprint*, arXiv:2505.00047.

Alissa Wilkinson. 2023. Hollywood's writers are on strike. here's why that matters.

Justin Wong, Yury Orlovskiy, Michael Luo, Sanjit A Seshia, and Joseph E Gonzalez. 2024. Simplestrat: Diversifying language model generation with stratification. *arXiv preprint arXiv:2410.09038*.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2024. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *arXiv preprint arXiv:2501.00273*.

Min-Hsuan Yeh, Leitian Tao, Jeffrey Wang, Xuefeng Du, and Yixuan Li. 2024. How reliable is human feedback for aligning large language models? *arXiv preprint arXiv:2410.01957*.

Yuhua Yu, Lindsay Krebs, Mark Beeman, and Vicky T Lai. 2024. Exploring how generating metaphor via insight versus analysis affects metaphor quality and learning outcomes. *Cognitive science*, 48(8):e13488.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. Noveltybench: Evaluating creativity and diversity in language models. *arXiv preprint arXiv:2504.05228*.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. 2024. Forcing diffuse distributions out of language models. *arXiv preprint arXiv:2404.10859*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, and 1 others. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. 2025. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms. *arXiv preprint arXiv:2502.20356*.

Aleksandra Zielińska, Peter Organisciak, Denis Dumas, and Maciej Karwowski. 2023. Lost in translation? not for large language models: Automated divergent thinking scoring performance translates to non-english contexts. *Thinking Skills and Creativity*, 50:101414.

A MuCE Dataset

We compiled data by means of crowdsourcing and data mining of the open-source data sharing platform OSF. We crowdsourced from the global creativity research community by means of direct requests and posts on academic listservs. In our call for data-sharing, we requested data relating to any creativity responses that were provided by human participants and scored for creativity by human

raters. We specifically requested that the datasets include scores from each rater, rather than composite creativity scores, to determine rating data quality for each submission. As part of our inclusion criteria, we further requested that researchers provide information relating to: (a) the creativity task, (b) the item associated with each response, (c) the construct that was rated, and (d) the language of the task. We further asked researchers to provide a statement on whether they agreed to making their data open-source. In terms of data mining through the OSF platform, we first searched through a series of relevant keywords (e.g., "creativity task", "originality score"). We only retained sub-datasets from credible sources, which were associated with a citable peer-reviewed article, and which included all the required data relating to our inclusion criteria.

After removing responses that didn't meet our inclusion criteria, our dataset amounted to 321,572 human-rated and language-based creativity responses. The dataset was thus cleaned by standardizing the naming for each variable except for the responses. We then removed responses for having been rated by fewer than 2 human judges. Duplicate responses were also removed, by retaining a single exemplar for responses that appeared twice within a specific item and task.

To enhance the reliability of human creativity ratings across the numerous datasets, we optimized the selection of raters by applying a meta-heuristic algorithm. Specifically, we applied a Genetic Algorithm (Schroeders et al., 2016). The GA operates through iterative selection, crossover, and mutation processes, mirroring the principles of natural selection, and in our case to identify the optimal subsets of raters for each dataset. In each iteration, candidate solutions—that is, combinations of raters—were evaluated based on a predefined fitness function that prioritized the maximization of empirical reliability (rxx) within a graded response model (GRM) and hence in line to judge response theory. For sub-datasets involving decimal-based scales, individual ratings were rounded to the nearest integer value (rounding up if containing a decimal = .5) to meet the requirements of the GRM.

Rater subsets demonstrating superior reliability were selected, recombined, and modified through random perturbations to prevent premature convergence to suboptimal solutions. This approach ensured that the selected raters provided consistent and informative judgments while reducing noise introduced by inconsistent or uninformative ratings. By automating the selection process through GA, we opted for maximal comparability in the selection process across datasets. Previous research has demonstrated the utility of GA in psychometric optimization tasks, particularly in balancing brevity and measurement precision while maintaining construct validity. In the present study, GA facilitated a systematic and data-driven refinement of rater selection, arguably enhancing the overall quality of creativity ratings.

After dropping uninformative raters in each sub-dataset, we again removed any rows containing less than 2 ratings due to rater removal. Afterwards, we used the new rater subsets per dataset and computed factor scores for each given response that were used as creativity scores. We calculated factor scores via a GRM model, ran separately over each sub-dataset, to derive a single creativity score for each response. Finally, we applied min-max scaling on each sub-dataset to transform ratings into a range of 10 to 50, with intervals of 1. This step was applied to ensure that ratings would only constitute a single token in length, to lessen the burden of predicting multi-token labels by the LLMs.

We then withheld all responses in the Spanish language from our final dataset and assigned them to an out-of-distribution-language (OOD-l) set. Responses from the OOD-l set were not included in the training data of MUCE, allowing us to test whether the model could generalize to creative responses in an unseen language. We selected Spanish as it would allow for a fair test of generalizability given: (1) Spanish tends to be a high-resource language within the pre-training of modern LLMs, (2) it is similar to other Latin-root languages in our training data (e.g., Italian), (3) responses in Spanish spanned multiple creativity tasks, and (4) the language spanned a limited number of responses in our total dataset. We further withheld all responses from two highly-naturalistic tasks, the Poem and Alternative Title Generation, and assigned these to an out-of-distribution task (OOD-t) set. We selected these tasks as they made up a limited portion of the total dataset and would provide a test of MUCE's performance on unseen naturalistic creativity tasks.

We then randomly selected items within each task and assigned them to an out-of-distribution item (OOD-i) set. We identified candidate items that corresponded to 5% or less of the responses within a task. Then, for tasks that contained 20 or

more total items, we randomly assigned 2 of these items to our OOD-i set. For tasks that contained fewer than 20 total items, we instead randomly assigned 1 of these items to the OOD-i set. Finally, we split the remaining responses in our dataset into training, validation, and out-of-distribution responses (OOD-r) sets according to an 80/10/10 split. We grouped responses into unique combinations of sub-dataset, task, language, item, and rating label, then randomly assigned responses within each combination to each of the sets, ensuring an equal representation of responses associated with each of these variables within the training, validation, and OOD-r sets. Table 1 contains the final dataset statistics for MUCE. Tables 8 and 9 contain the descriptions and data statistics for each task in MUCE. Tables 10, 11, 12, 13, and 14 list some example prompts and low-rated and high-rated responses for each task from MUCE.

B SFT and Preference Datasets

Past work has shown that data quality is one of the main factors behind preference model performance (Liu et al., 2024; Deng et al., 2025; Wang et al., 2024a). In particular, the margin in the score (i.e. reward margin) between the preferred and rejected response may influence the performance of the model, since training pairs with smaller margins are likely to contain annotation noise and be more difficult to learn. We experiment with different reward margins and choose a margin of 5 for the final experiments as it showed a balance between mitigating annotator noise and creating a dataset with nuanced preferences. Additionally, to ensure a high-quality preference dataset, first we filter the base MUCE dataset and select only the samples that have a full agreement from all annotators. Then we filter out all samples that have a rating below 20 and limit the number of pairings between samples to 10. This results in a final preference training dataset of $42,058$ samples (MUCE-PREF). We also create a high-quality instruction-tuning dataset from MUCE-PREF by pairing the prompts with all preferred responses that have a rating above 30 resulting in a dataset of $5,285$ samples (MUCE-SFT). Tables 3 and 4 contain the statistics for these datasets.

C Training

We follow a training setup similar to Chung et al. (2025) and use `Llama-3.1-8B-Instruct` and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023) as our base models. Using these models, we train an SFT, DPO and several CRPO models. We train all models using parameter-efficient tuning with LoRA using a rank of 128 and an alpha of 256 (Hu et al., 2022). All training was done using HuggingFace TRL library[7] with Accelerate (Gugger et al., 2022) and DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) on `NVIDIA A100` GPUs with gradient checkpointing.

SFT model is trained on the MUCE-SFT dataset for a single epoch with a batch size of 2 per GPU using a gradient accumulation size of 4 and context size of 1024. We use a cosine scheduler with a half-cycle warmup and maximum learning rate of $3e-5$. Final model achieves 85% mean token accuracy on the validation set.

DPO and CRPO models are trained using the SFT model as a base on our MUCE-PREF dataset for a single epoch with a batch size of 8 per GPU using a gradient accumulation size of 8 and context size of 1024. We use a linear scheduler with a learning rate of $5e-6$. All final models achieve over 82% reward accuracy on the validation set.

D Evaluation

For each prompt in our held-out evaluation suite, we generate a total of 16 responses for every model by sampling 4 responses for each of the following four decoding setups that induce high randomness using various sampling techniques (Fan et al., 2018; Holtzman et al., 2019):

1. $temperature = 0.7, top_p = 0.95$
2. $temperature = 0.9, top_p = 0.99$
3. $temperature = 0.7, top_k = 50$
4. $temperature = 0.8, top_p = 0.97$

Moreover, as the existing instruction-tuned LLMs tend to produce verbose outputs (Saito et al., 2023), in order to minimize the length bias, we add further instructions in the prompt, constraining the output length in terms of the number of sentences and words. We compute the constraint values based on the median number of words and sentences of responses per task from our training dataset. Table 5 lists an example evaluation prompt for each task. Table 7 lists an example response from all models to a single prompt.

[7] https://huggingface.co/docs/trl/en/index

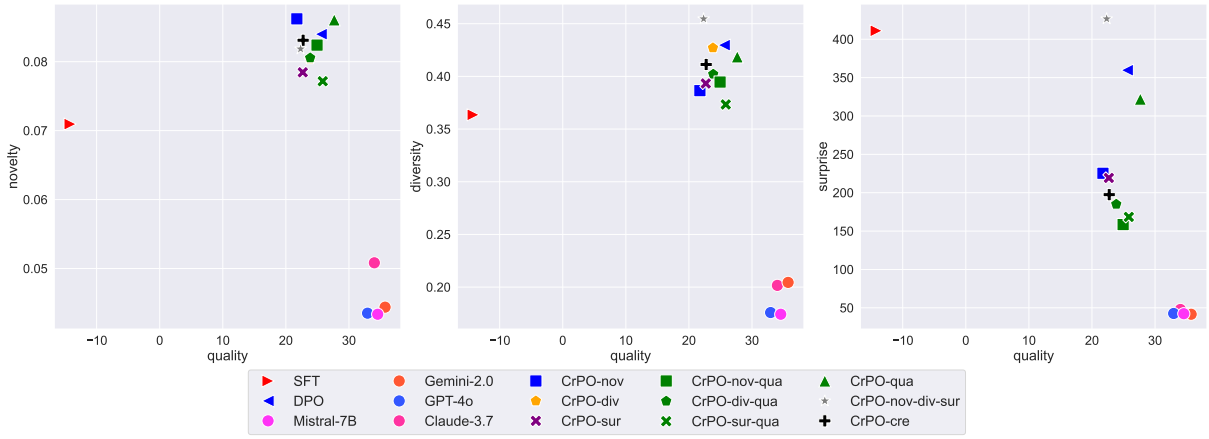

Figure 6: Results on held-out evaluation suite from MUCE across all baselines and our models using `Mistral-7B-Instruct-v0.3` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks.

	Total	Train	Dev	Test	OOD-i	OOD-l	OOD-t
# samples	245,030	183,973	23,254	22,419	6,253	4,719	4,412
# tasks	25	23	23	23	9	3	2
# languages	11	10	10	10	5	1	3
# prompts	587	503	502	502	12	52	21

Table 1: Detailed statistics for each split of MUCE.

Human Evaluation Instructions

In this study, you will be presented with two responses to a creative task. Your job is to select the response that you believe is the most creative. Please base your judgment only on the creativity of the ideas—not on how long or detailed the response is. A shorter response can be more creative than a longer one, and vice versa. Focus on how original, unique, and innovative the idea feels to you. There are no right or wrong answers—we're interested in your opinion.

Figure 7: Rater instructions for the human evaluation.

D.1 Human Evaluation

Since we have multiple model responses per prompt, instead of randomly choosing a response, for each prompt, we choose top 4 model responses measured by the overall automated creativity score which we define as the sum of normalized novelty, diversity, surprise and quality scores. This setup ensures that models are compared to each other with their best outputs. We recruited 15 participants on Prolific[8] to complete the study, requiring that they reside in the U.S. and have an approval rating of at least 90%. Ethics board approval was received from the Pennsylvania State University IRB for this study. We provided participants with a definition of creativity, and instructed them not to focus on the length or detail of the response when rating. Figure 7 lists the instructions given to raters for evaluating creativity. We additionally included a comprehension check where participants were quizzed about the task instructions, to help catch careless participants. Raters who failed this check were excluded from further analysis. All raters were compensated adequately with at least a minimum payment of 9$ per hour. Final win rates are calculated for each response pair based on the majority vote across participants. The inter-rater agreement computed using Krippendorff's alpha was 0.463, indicating a moderate agreement.

D.2 NOVELTYBENCH Evaluation

NOVELTYBENCH is a recently introduced benchmark to measure how well language models can generate novel and high-quality answers to user requests involving subjectivity, randomness, and creativity (Zhang et al., 2025). We use a 100-sample

[8] https://www.prolific.com/

Model	novelty	diversity	surprise	quality
GPT-4o (std. prompting)	0.04	0.18	42.63	**32.94**
GPT-4o (Brainstorm & Select)	0.05	0.28	14.26	22.01
CRPO-nov	**0.10**	0.43	143.98	27.46
CRPO-div	**0.10**	**0.49**	204.48	28.87
CRPO-sur	**0.10**	**0.49**	**590.65**	17.91
CRPO-nov-div-sur	**0.10**	0.47	249.31	27.07
CRPO-cre	**0.10**	0.44	203.17	28.55

Table 2: AUT evaluation results comparing a creative prompting strategy such as Brainstorm & Select, to our approach with standard prompting. Best performance is **bolded** for each metric.

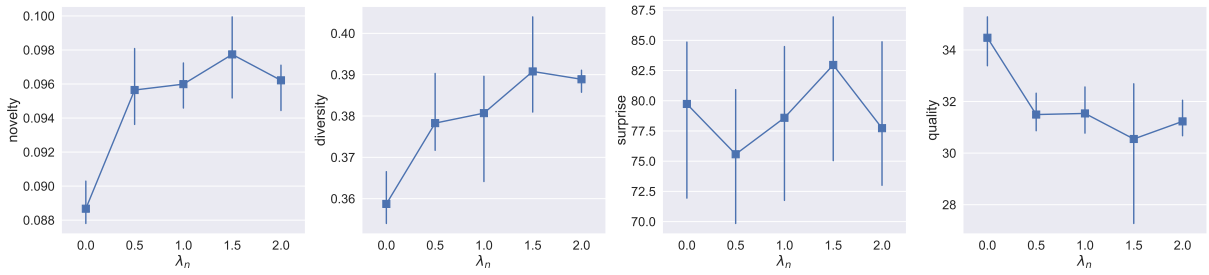

Figure 8: Effect of novelty injection weight on all creativity dimensions. Results are averaged across three seed runs.

subset of their benchmark that is manually curated by the authors and contains four distinct categories where diversity and novelty are expected:

- **Randomness:** prompts that involve randomizing over a set of options. Example: *Roll a make-believe 20-sided die.*
- **Factual Knowledge:** prompts that request underspecified factual information, which allow many valid answers. Example: *List a capital city in Africa.*
- **Creative Writing:** prompts that involve generating a creative form of text, including poetry, and story-writing. Example: *Tell me a riddle.*
- **Subjectivity:** prompts that request subjective answers or opinions. Example: *What's the best car to get in 2023?*

Additionally, the paper proposes new metrics to measure novelty and quality (i.e. utility) that are different than ours. To compute novelty, they propose a method that learns to partition the output space into equivalence classes from human annotations. Each class represents one unique generation that is roughly equivalent to the others in the same class and different from the generations in other classes. They consider a functional equivalence that defines two generations to be different if and only if a user who has seen one generation would likely benefit from seeing the other. To this end, the authors annotated 1,100 pairs of generations conditioned on prompts from NOVELTYBENCH sampled from a diverse set of models. From these annotated pairs, they used 1,000 for training and fine-tuned a deberta-v3-large model (He et al., 2023) to predict binary functional equivalence between two generations. With the equivalence classifier, they partition the output space into equivalence classes. Then they define the novelty as the $distinct_k$ metric that is the number of equivalence classes in a partition of k sample generations from a language model:

$$distinct_k := |\{c_i | i \in [k]\}| \tag{8}$$

To compute quality, they consider a model of user behavior that describes how users interact with and consume language model generations. They assume that the user has a patience level $p \in [0, 1]$: after observing each additional generation, they have a probability p of requesting an additional generation from the language model and observing the next generation, and a probability $1 - p$ of stopping interacting with the model. Then they compute the quality of a sequence of generations

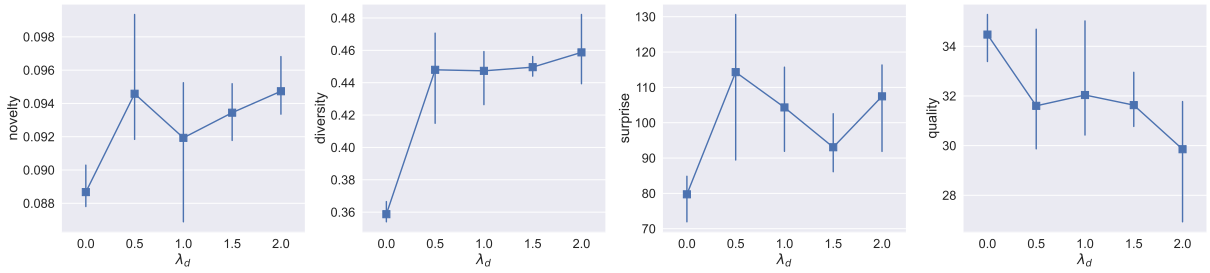

Figure 9: Effect of diversity injection weight on all creativity dimensions. Results are averaged across three seed runs.

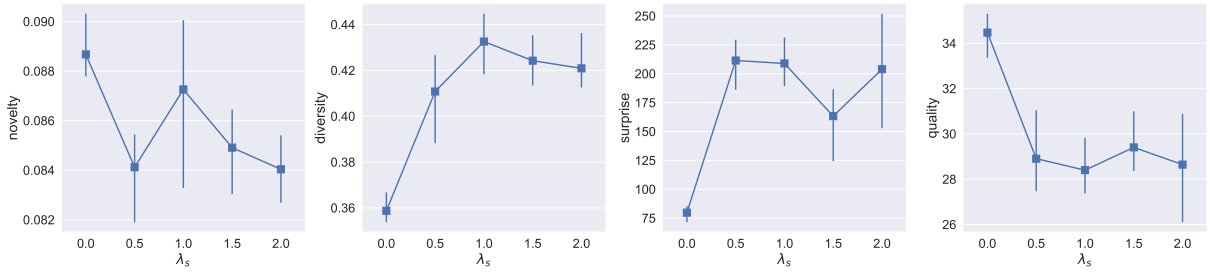

Figure 10: Effect of surprise injection weight on all creativity dimensions. Results are averaged across three seed runs.

as the cumulative utility:

$$utility_k := \frac{1-p}{1-p^k} \sum_{i=1}^{k} p^{i-1} \cdot \mathbb{1}[c_i \neq c_j, \forall j < i] \cdot u_i \tag{9}$$

To compute the utility of individual generations, they also use the `Skywork-Reward-Gemma-2-27B-v0.2` (Liu et al., 2024) model.

To benchmark our models, we follow their recommended setup for evaluation. In particular, we set the number of generations to 10 per model and the patience level to 0.8 and use their trained classifier for output space partition.

D.3 Additional Baselines

We consider additional baseline comparisons using creative prompting and decoding strategies to further show the effectiveness of our approach.

D.3.1 Creative Prompting

We choose an approach based on brainstorming and selecting (Summers-Stay et al., 2023) that has shown improvement on creativity scores in the Alternative Uses of Objects Task (AUT). In this method, an iterative prompting strategy is employed by asking the model first to generate a response, then brainstorm about the advantages and drawbacks of its response, and then evaluate the novelty and utility of its response, and propose different uses if needed. We follow the prompts suggested in Summers-Stay et al. (2023) and evaluate one of our best-performing baselines, namely, GPT-4o model on the AUT task with a held-out set of 25 objects where for each object we generate 16 different responses using the same evaluation setup mentioned above. We also evaluate our CRPO models on the same held-out test suite, but with standard single-turn prompting.

D.3.2 Creative Decoding

Our main evaluation setup provides comparisons to popular decoding strategies such as top-p, top-k and temperature sampling that are often used to increase the creativity of language model outputs. However, these approaches are not designed to directly improve output creativity. Hence, to further show the effectiveness of our approach, we consider a recent decoding strategy called *min-p sampling* that promotes output creativity directly (Nguyen et al., 2024). We follow the paper's recommended setup for best performance (min-p values 0.05, 0.1 with high-temperature values 1.0, 1.5) on creative tasks and evaluate the base model `Llama-3.1-8B` with this decoding setup on our multi-task evaluation suite.

Task	# prompts	# samples
Real-Life Creative Problem Solving	8	5,601
Question Asking	5	314
Malevolent Problems	21	424
Metaphors	51	675
Alternate Uses of Objects Task	11	4,388
Design Solutions	10	1,366
Essays	1	174
Stories	7	1,498
Consequences	5	10,865
Experiment Design	7	5,640
Hypothesis Generation	6	5,260
Research Questions	5	5,832
Associations	5	21
Total	**142**	**42,058**

Table 3: MUCE-PREF training dataset details.

Task	# prompts	# samples
Real-Life Creative Problem Solving	8	642
Question Asking	6	58
Malevolent Problems	22	82
Metaphors	60	158
Alternate Uses of Objects Task	11	855
Design Solutions	12	150
Essays	1	23
Stories	7	256
Instances of Common Concepts	4	10
Consequences	5	1,315
Experiment Design	7	573
Hypothesis Generation	6	548
Research Questions	5	587
Associations	7	28
Total	**161**	**5,285**

Table 4: MUCE-SFT training dataset details.

Task	Prompt
Real-Life Creative Problem Solving	"Come up with an original and creative solution for the following real-world problem: Clara, a junior pre-med student, is working part-time and taking a 15 hour credit load at school. ...<skipped>... Please limit your response to 4 sentences and at most 75 words."
Alternate Uses of Objects	"Come up with an original and creative use for the following object: rope. Please limit your response to 1 sentence and at most 17 words."
Design Solutions	"Come up with an original and creative solution to reduce the amount of litter in public spaces and promote waste reduction and recycling. Please limit your response to 2 sentences and at most 36 words."
Hypothesis Generation	"Come up with an original and creative scientific hypothesis for the following scenario: You notice that dogs seem to like one of your friends, but cats seem to like another friend. What hypotheses do you have about why that is? Please limit your response to 1 sentence and at most 22 words."
Metaphors	"Come up with an original and creative metaphoric equivalent for the concept described below: Stomata are tiny openings or pores found on the underside of a plant leaf. They are used for gas exchange, enabling the intake of carbon dioxide and release of oxygen.. Please limit your response to 1 sentence and at most 10 words."
Poems	"Come up with an original and creative poem about the following concept: choice. Please limit your response to 5 sentences and at most 150 words."
Sentence Completion	"Finish the sentence with an original and creative ending: When I got on the school bus....Please respond in one sentence."

Table 5: MUCE dataset held-out item and task evaluation prompts.

Model	novelty	diversity	surprise	quality
`Llama-3.1-8B`	0.05	0.21	46.28	31.59
`Llama-3.1-8B` (min-p)	0.06	0.22	51.42	31.31
CRPO-nov	**0.10**	0.36	71.75	32.56
CRPO-div	0.09	**0.46**	115.73	30.42
CRPO-sur	0.08	0.44	**230.91**	27.39
CRPO-nov-div-sur	0.09	0.43	120.07	29.24
CRPO-cre	0.09	0.39	96.92	**33.51**

Table 6: Evaluation results comparing a creative decoding strategy min-p sampling, to our approach using `Llama-3.1-8B` as a base model. Best performance is **bolded** for each metric.

Model	Response
Human	"played hide-and-seek by forming into different objects and trying not to be found."
Llama-3.1-8B-Instruct	"At a sleepover, we transformed our host's basement into a mystical 'Dreamscape' where we wove a tapestry of secrets, stories, and whispered promises that only the darkness could keep."
Gemini-2.0-Flash	"At a sleepover we...built a pillow fort so magnificent, it accidentally opened a portal to a realm where dreams manifested as sentient, fluffy creatures demanding to be tickled."
Claude-3.7-Sonnet	"At a sleepover we constructed an elaborate blanket fortress complete with secret passages, flashlight constellations on the ceiling, and an unspoken pact to guard each other's midnight confessions until the end of time."
GPT-4o	"At a sleepover we transformed the living room into a magical fort of pillows and fairy lights, where we whispered secrets and told stories until the first light of dawn."
SFT	"We played a game of Twister where we put out our phones to see who was the best twister."
DPO	"conducted an experiment to see if a person's personality is changed with an altered state of consciousness."
CRPO-nov	"tried to break the record for most consecutive hours without speaking and we discovered we could all hear each other's thoughts."
CRPO-nov-qua	"Participated in an experiment where we tested the effects of sleep deprivation on the human mind."
CRPO-div	"construct a space shuttle that takes us to the moon and from there we can launch our dream rockets."
CRPO-div-qua	"recreated the conditions of a 19th century underground railroad and had to map out our escape to Canada."
CRPO-sur	"Operate on each other to implant a permanent adrenaline gland."
CRPO-sur-qua	"created an underwater laboratory within our inflatable pool to collect the evidence we found of alien life."
CRPO-qua	"began to master the art of telekinesis by competitively tossing each other's pillows across the room."
CRPO-nov-div-sur	"Built a rollercoaster out of air mattresses and then did a hot-wheel car-launch into the trenches and caught the crash on GoPro cameras."
CRPO-cre	"Created an experiment to see if our dreams could be manipulated and transfer to one another."

Table 7: Example model (and gold human) responses to the prompt "At a sleepover we ..." from the *Sentence Completion* task.

Task	Description	Dataset Sources	# prompts	# samples
Real-Life Creative Problem Solving	Produce solutions for everyday problems	(Luchini et al., 2025; Kapoor et al., 2024; Saretzki et al., 2024; Perchtold-Stefan et al., 2024)	28	33,340
Alternate Titles Generation	Produce alternative titles for widely known books or movies	(Agnoli et al., 2016)	6	2,986
Question Asking	Produce questions about everyday objects	(Raz et al., 2024)	6	3,566
Poems	Produce poems about everyday concepts	(Fan et al., 2023; Chaudhuri et al., 2025; He et al., 2022)	15	2,580
Design Solutions	Produce solutions to real-world design problems	(Luchini et al., 2025; DiStefano et al., 2025)	20	10,818
Combining Objects	Produce combinations of everyday objects to achieve a goal	(Weiss et al., 2023)	13	4,494
Plot Titles Generation	Produce titles for story plots	(Weiss et al., 2023; Goecke et al., 2024b; Weiss et al., 2024)	6	1,832
Instances of Common Concepts	Produce instances related to everyday adjectives	(Organisciak et al., 2023)	8	2,474
Experiment Design	Produce experiment designs to test scientific hypotheses	(Beaty et al., 2024; Goecke et al., 2024a)	7	4,893
Associations	Produce word associations	(Beaty and Johnson, 2021)	8	1,004
Emotional Trials	Produce feelings one might have in a given situation	(Weiss et al., 2023)	6	732
Invent Nicknames	Produce nicknames for everyday concepts and objects	(Weiss et al., 2023)	6	613
Situation Redescription	Produce redescriptions of negative situations into positive situations	(Weiss et al., 2023)	6	826
Alternate Uses of Objects Task	Produce alternate uses for everyday objects	(Patterson et al., 2023; Zielińska et al., 2023; Organisciak et al., 2023)	211	88,155
Stories	Produce short stories from three word prompts	(Luchini et al., 2025; Agnoli et al., 2016; Fan et al., 2023; He et al., 2022)	10	2,757

Table 8: MUCE dataset details broken down by task (Part 1).

Task	Description	Dataset Sources	# prompts	# samples
Malevolent Problems	Produce ideas on how to take revenge on or sabotage a wrongdoer	(Perchtold-Stefan et al., 2023; Kapoor et al., 2024; Perchtold-Stefan et al., 2024)	36	16,536
Metaphors	Produce metaphors to describe scenarios	(DiStefano et al., 2024; Yu et al., 2024)	136	13,210
Essays	Produce essays on a topic	(Cotter et al., 2016)	1	821
Consequences	Produce possible consequences to scenarios	(Weiss et al., 2024, 2023; Goecke et al., 2024b)	20	24,874
Sentence Completion	Produce endings to incomplete sentences	(Organisciak et al., 2023)	8	2,629
Hypothesis Generation	Produce scientific hypotheses for specific observations	(Beaty et al., 2024; Goecke et al., 2024a)	9	18,455
Research Questions	Produce research questions relating to scenarios	(Beaty et al., 2024; Goecke et al., 2024a)	5	5,161
Composites	Produce composite words from a prompt word	(Weiss et al., 2023)	7	667
Evoking Emotional Responses from People	Produce ways to evoke emotional responses in people as a TV producer	(Weiss et al., 2023)	6	789
Emotions in Everyday Situations	Produce emotions you might feel in response to everyday situations	(Weiss et al., 2023)	6	818

Table 9: MUCE dataset details broken down by task (Part 2).

Task	Example prompt	Example low rating response	Example high rating response
Real-Life Creative Problem Solving	"Becky is a college student who works part-time at Mark's Pizzeria. Mark, the owner of the restaurant, has treated Becky very well. He gave her a job that she needs to help pay her rent when no other business would employ her because she was arrested for shoplifting three years ago. Mark also lets Becky work around her school schedule, and has asked if she wants to be a shift manager in the summers. Becky's roommate Jim also works at the pizzeria, but Jim has been causing a lot of problems at work. He always avoids doing his job, treats customers rudely, and makes a lot of mistakes with orders. Jim recently began stealing food from the pizzeria. Two days ago the pizzeria was short- staffed, so Jim and Becky were the only employees left at closing time. Jim made 10 extra pizzas and took them home to a party he was hosting without paying for them. Becky feels like she needs to do something about Jim's behavior. However, Becky is hesitant to tell Mark about Jim because Jim is a good friend to Becky. Becky also needs Jim to have a job so he can pay his portion of their rent. Becky does not know what to do.."	"Morally the right thing for Becky to do would be to tell her boss. However, to be a good friend would to be not to tell on Jim. The only creative solution to this problem would to be to try and talk to Jim one on one. Give Jim the decision of whether or nt he wants Becky to inform their boss of what he has been doing. As a friend he should understand where Becky is coming from and want to take the strain off her."	"Becky should first discuss this with Jim, and tell him that he needs to either pay for the pizzas or he needs to go to the boss, and admit what he has done. He will get caught in the end because eventually the ingredients will be missed. The boss may unerstand, and perhaps will allow him to work off the pizzas somehow. Maybe he could help out cleaning up around the restaurant. If Jim will not tell his boss Becky should tell him. She wouldn't necessarily have to come right out and tell on her coworker she could come up with a way for the boss to catch him at it. If he does it once Jim will more than likely do it again. She could tell the boss to check on the inventory. She could have other people who might have been at the party come tell her boss bout it. If all of that fails, she should just tell Mark about Jim stealing the pizzas."

Table 10: MUCE dataset examples (Part 1).

Task	Example prompt	Example low rating response	Example high rating response
Question Asking	"pencil"	"How big is it?"	"How many great ideas have started with a pencil?"
Poems	"childhood"	"Twinkle, Twinkle little star....ect"	"Red Rover, Red Rover Is my childhood over? I don't feel quite grown up I still laugh at "I CUP" I play slide with my sister and still call my fourth grade teacher "mister" I suppose, even still, my childhood is over even if I can still play red rover red rover"
Design Solutions	"Develop as many design ideas as you can to reduce air pollution in cities."	"Walk"	"use 3d printing as an innivating way of building houses as it reduces labour and"
Combining Objects	"Paint sign"	"paper, ballpoint pen"	"beetroot juice, quark cheese"
Plot Titles Generation	"Now spoke"	"A completely normal everyday life"	"VR glasses charger defective"
Instances of Common Concepts	"soft"	"something that is not hard"	"a futuristic ball that turns really fuzzy and comfy at places it gets contact to"
Experiment Design	"You think some animals have a sense of humor that humans don't usually understand. How could you test that hypothesis?"	"observe"	"tickle your dog to see how he acts when he's 'laughing.' then, observe your dog throughout the day and note when he is 'laughing.' you may begin to pick up on moments where he does things that are funny to him."
Associations	"expert"	"winner"	"ace"
Emotional Trials	"You have a date tonight, and once again your dress didn't get ready in time at the laundry."	"worried, afraid, sad"	"Anger, panic, anticipation"
Invent Nicknames	"plate"	"porcelain"	"Shrunken UFO"

Table 11: MUCE dataset examples (Part 2).

Task	Example prompt	Example low rating response	Example high rating response
Alternate Uses of Objects Task	“knife”	“weapon”	“make up "knife characters" and create a movie”
Stories	“petrol-diesel-pump”	“I needed to fuel my car before we could start the long drive. I drove to the petrol station. i went to the pump and fuel my car with diesel. new i was ready for the task ahead”	“Manly Merde was a truck driver looking for trouble. He pulled into the Casino in the back where the drivers go. He took a swig of whisky and walked to the petrol station, grabbed the pump and spurt diesel into the air like hydrocarbon fountain. He let out a big belly laugh and screamed, "Let the revolution begin!" And that is how the trucker wars started.”
Malevolent Problems	“Your professor in class announces an award for the person who comes up with the best solution for a project. By chance, another student leaves their notebook behind in class. You read their ideas and believe that they are the best. You decide to turn them in as your own; however you know that if the other student submits the same solution, there will be a problem.”	“I will not do the above”	“render their notebook unreadable by dropping water at the last moment”
Metaphors	“The hot tea is...”	“boiling”	“liquid fire”
Consequences	“What would be the result if society no longer used money, and instead traded goods and services?”	“Banks would be unnecessary.”	“People (especially couples) would stop fighting so much about financial issues”
Sentence Completion	“It started raining and...”	“I got wet”	“because I was covered in oil, I began to levitate, and all the witnesses called me the next coming of some sort of goddess.”

Table 12: MUCE dataset examples (Part 3).

Task	Example prompt	Example low rating response	Example high rating response
Hypothesis Generation	"On a field trip, you drive past a massive field with hundreds of large holes visible as far as the eye can see. What hypotheses do you have about what purpose the holes may serve?"	"the holes resulted over time and nature"	"the holes are for animals giving birth."
Essays	"dream project"	"I don't really know what carreer path I want to follow. I just want a job where I can help people and get a good pay check so I can support myt future endevors. I want to do something that no one has ever done before in a way no one has ever seen. I want to inspire a generation to work on a better future for everybody. I guess what I really want is to be remembered as an icon. i want to be someone that people look up to."	"I want to go into forensic science when I gradutate. Therefore, my dream project is to discover the perfect device that can help solve every crime scene. This device would be able to analyize the crime scene and tell us exactly how many people died and how they died. It would then collect evidence samples such as blood. Next, it would use what the information it found at the crime scene to help make up questions the dectectives would ask the suspects. It would use it's technology to come up with questions that only the murderer could answer. Later on, back at the lab, it would help discover whose blood the samples belonged to. In the end, the only human power that would be needed was someone to arrest the convicted person and the people to help clean up the crime scene."

Table 13: MUCE dataset examples (Part 4).

Task	Example prompt	Example low rating response	Example high rating response
Situation Redescription	"You notice how your colleague first treats another employee very kindly and then shortly afterwards starts talking negatively behind his back"	"It would be nice if you were older"	"I'll talk to them. Then I'll have to work less"
Alternate Titles Generation	"The Betrothed"	"renzo and lucia"	"Plague, Honor and Love in Baroque Brianza"
Research Questions	"You travel to a jungle that contains no human life and is completely unknown to the scientific community. What scientific questions could you ask about this jungle?"	"How many people will come with me?"	"Do these species share a common characterisitic that humans don't have?"
Composites	"jitters"	"Exam jitters"	"Easter bunny missing jitters"
Evoking Emotional Responses from People	"Describe how you would make people look down on others"	"I will always scream loudly"	"I would divide the audience into two groups and give one group a rubber glove as headgear and the other group a tiara or crown made of real gold.'
Emotions in Everyday Situations	"You're at work. A glance at the clock tells you that you're about to finish work and start your long-awaited weekend."	"I feel happy"	"I feel sorry for my desk chair, which is unused over the weekend and stands alone in the office."

Table 14: MUCE dataset examples (Part 5).