HUMANE Paper Review

Creative Preference Optimization

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

[1]EPFL, [2]Università della Svizzera Italiana, [3]Pennsylvania State University
mahammad.ismayilzada@epfl.ch

숭실대학교 문화콘텐츠학과, 석사과정생 이다현

preprint

2025.06.13

Background

- Whether LLMs exhibit true human-like creativity remains unclear
- Some studies find LLMs more creative than humans, while others disagree

- Limitations of Current LLM Creativity
  - LLMs often lack novelty and surprise
  - generate less diverse content compared to humans
- Existing Solutions & Gaps
  - Prior methods target <u>diversity</u> or <u>single tasks</u> only
  - Creativity is multi-dimensional(novelty, surprise, quality, and diversity)
    → need methods that optimize multiple creativity dimensions

Background

- Propose a new method
  → directly optimize creativity in LLM outputs using preference learning
  - Prior approaches mainly relied on black-box prompting or decoding
  - Recent studies also raise concerns that preference alignment reduces output diversity

Contributions

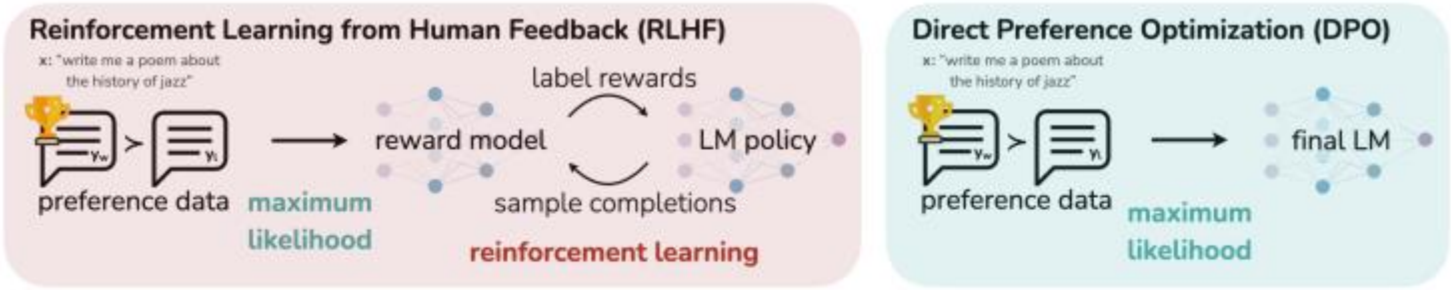
- Proposed method: Creative Preference Optimization (CRPO)
  - Propose Creative Preference Optimization (CRPO),
    extending DPO to inject multiple creativity signals
  - CRPO integrates novelty, diversity, surprise, and quality into the training objective
  - Tested on MUCE, a new multitask, multilingual dataset with human preference labels
  - CRPO outperforms baseline methods including SFT, DPO, and GPT-4o in creative generation

CRPO

- What is creativity?
  - Creativity is defined by three core criteria: Novelty, Quality, and Surprise (Simonton, 2012; Boden, 2004)
  - Additionally, diversity across individuals is a hallmark of creative output
- Problem with Existing Methods
  - Standard Preference Optimization (e.g., DPO) may reduce output diversity
  - Model is trained to focus only on preferred responses, even if they are not very creative

DPO

- method for aligning large language models (LLMs) directly with human preferences, without requiring reinforcement learning or a reward model

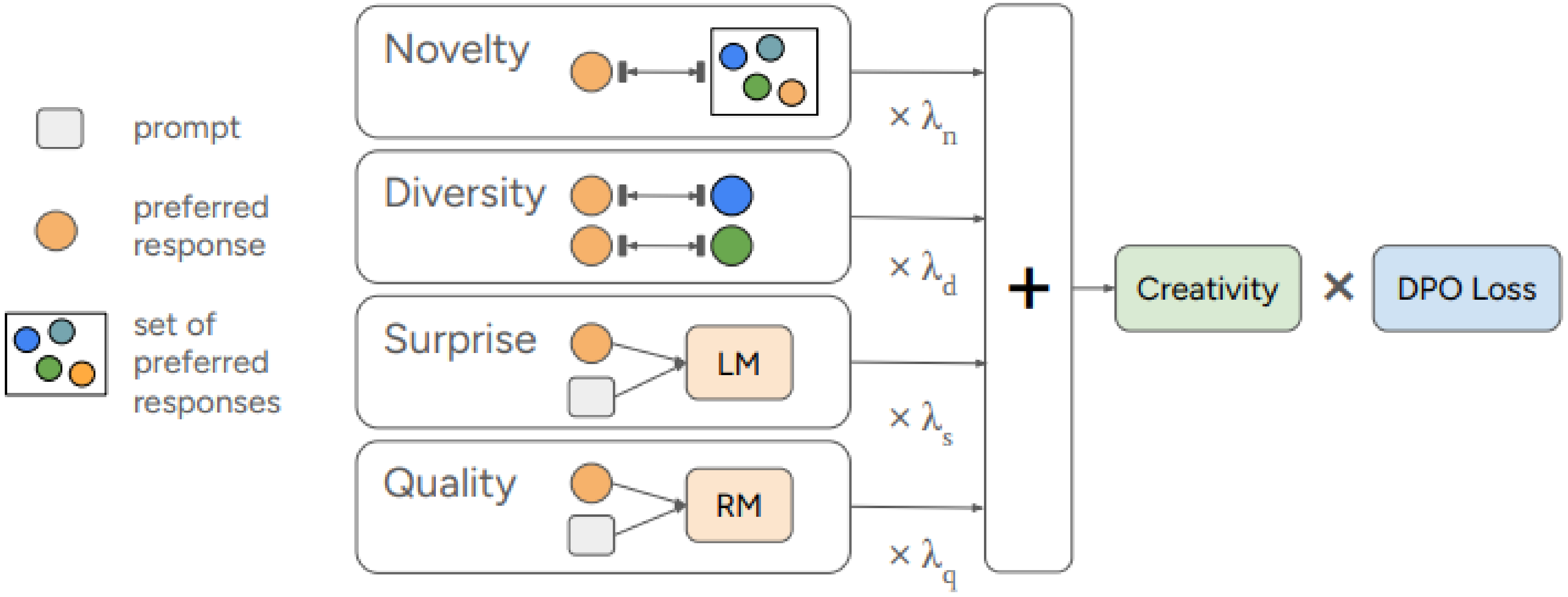


$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

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

CRPO

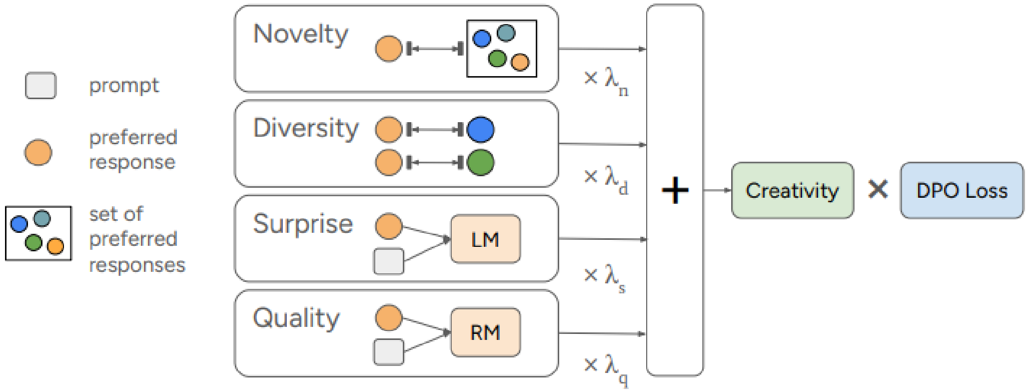
- Extend DPO to support multiple creativity dimensions
- $L_{CRPO} = -\mathbb{E}_{(x,y^w,y^l)\in D}\left[\left(\lambda_d\delta_w + \lambda_n\nu_w + \lambda_s\xi_w + \lambda_q\gamma_w\right) l_{DPO}\right]$



CRPO

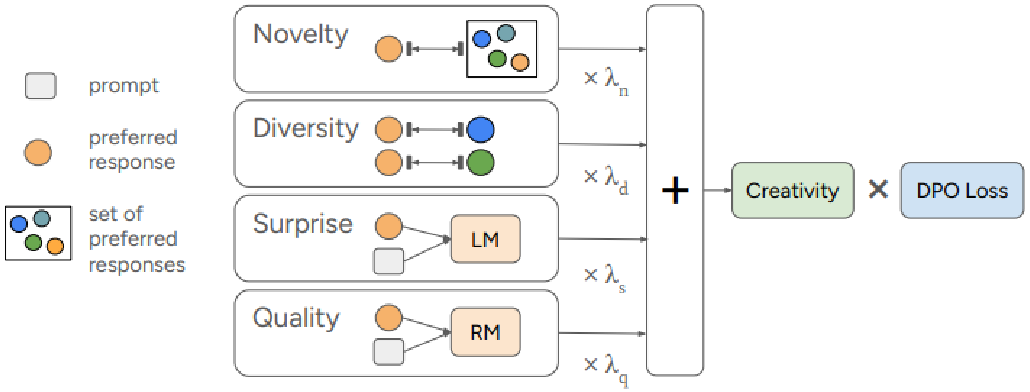
- Diversity
  - $\delta_w = \frac{1}{|Y_x| - 1} \cdot \sum_{y_i \in Y_x \backslash y_w} semdis(y_w, y_i)$

$$semdis(\cdot,\cdot) = 1 - cos\_sim(\cdot,\cdot)$$



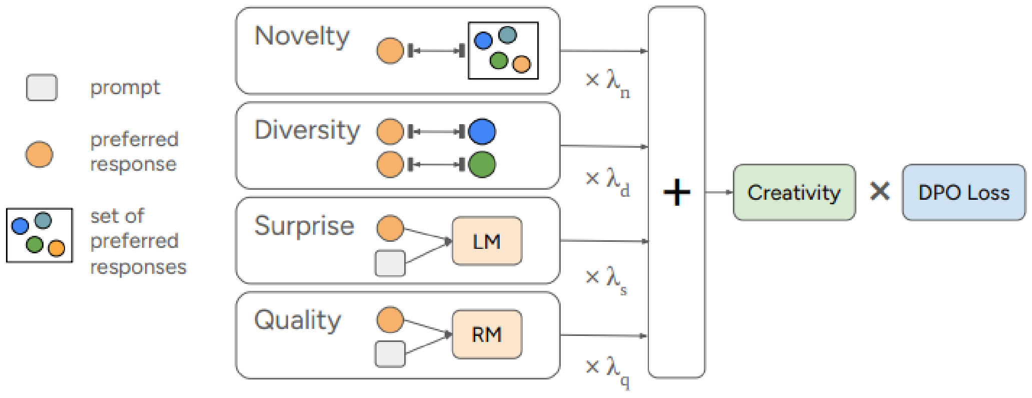
CRPO

- Diversity
  - $\delta_w = \frac{1}{|Y_x| - 1} \cdot \sum_{y_i \in Y_x \setminus y_w} semdis(y_w, y_i)$
- Novelty
  - DSI = Divergent Semantic Integration
  - $v_w = |DSI(y^w) - DSI(Y_x)|$
  - $DSI(T) = \frac{\sum_{i,j=1}^{|T|} semdis(T_i, T_j),\, i \neq j}{|T|}$



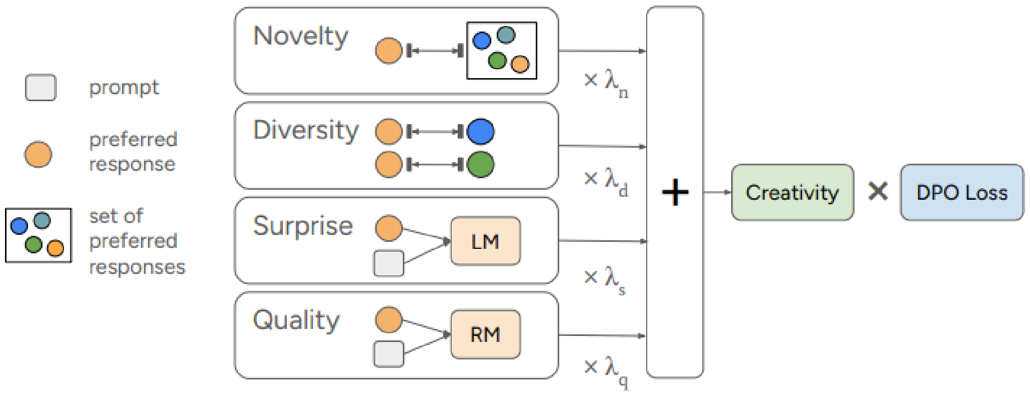
CRPO

- Diversity
  - $\delta_w = \frac{1}{|Y_x| - 1} \cdot \sum_{y_i \in Y_x \setminus y_w} semdis(y_w, y_i)$
- Novelty
  - $\nu_w = |DSI(y^w) - DSI(Y_x)|$
  - $DSI(T) = \frac{\sum_{i,j=1}^{|T|} semdis(T_i, T_j),\ i \neq j}{|T|}$
- Surprise
  - $\xi^w = 2^{-\log P_S(y_w|x)}$



CRPO

- Diversity
  - $\delta_w = \frac{1}{|Y_x| - 1} \cdot \sum_{y_i \in Y_x \backslash y_w} semdis(y_w, y_i)$
- Novelty
  - $\nu_w = |DSI(y^w) - DSI(Y_x)|$
  - $DSI(T) = \frac{\sum_{i,j=1}^{|T|} semdis(T_i, T_j),\, i \neq j}{|T|}$
- Surprise
  - $\xi^w = 2^{-\log P_S(y_w|x)}$
- Quality
  - $R: \gamma^w = R(y^w|x)$



Evaluation

- MUCE
  - Multitask Creativity Evaluation
  - Enables to test whether method truly generalize across a diverse range of creativity assessment

Evaluation

- MUCE
  - Multitask Creativity Evaluation
  - Enables to test whether method truly generalize across a diverse range of creativity assessment
  - Collected data from the global creativity research community, especially experts on human creativity tasks
  - 43% of data is unreleased, reducing LLM data leakage
  - Each response rated by 2–75 raters;
    applied Judge Response Theory (JRT) + genetic algorithm to filter noisy raters
  - Ratings aggregated using factor scores, then rescaled to 10–50

Evaluation

- SFT & Preference Datasets
  - English subset of MUCE
  - Preference dataset is built by pairing responses to the same prompt and selecting the one with a higher creativity score
  - From MUCE, curated
    :5,275 SFT samples (MUCE-SFT)
    42,058 preference pairs (MUCE-PREF)

Evaluation

- Training Setup
  - Base models: Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3
  - Training steps:
    1. SFT on MUCE-SFT (1 epoch)
    2. CRPO optimization using MUCE-PREF
- Creativity Signal Injection
  - For each preferred response, compute scores for:
    diversity, novelty, surprise, and quality
  - Scores are normalized to [0, 1] before injection

Evaluation

- Metric Computation Details
  - Prompt-level reference sets are used for novelty & diversity, following Chung et al. (2025)
  - Text embeddings computed using *jina-embeddings-v3* for semantic metrics
  - Surprise scores computed using *Gemma-2-27B* (instruction-tuned)
  - Quality measured via *Skywork-Reward-Gemma-27B-v0.2*, a top model on RewardBench

Evaluation

- Evaluation & Baselines
  - 224 held-out samples across 6 tasks + 2 unseen tasks (Poems, Sentence Completion)
  - For each prompt: 16 responses per model with varied decoding parameters
  - Evaluation dimensions: novelty, diversity, surprise, plus quality tradeoff via learned reward model

Evaluation

- Evaluation & Baselines
  - 224 held-out samples across 6 tasks + 2 unseen tasks (Poems, Sentence Completion)
  - For each prompt: 16 responses per model with varied decoding parameters
  - Evaluation dimensions: novelty, diversity, surprise, plus quality tradeoff via learned reward model
- Baselines:
  - Original Llama/Mistral
  - SFT on MUCE-SFT
  - Vanilla DPO on MUCE-PREF (no creativity)
  - GPT-4o, Claude 3.7, Gemini 2.0 Flash
- CRPO models:
  - For each creativity dimension alone
  - With & without quality injection
  - Full injection (CRPO-cre) and CRPO without quality (CRPO-nov-div-sur)
  - For simplicity, all λ values set to 1(모든 λ 값은 1로 고정하여 실험 단순화)

Result

- clear separation between existing instruction-tuned LLMs vs our Models

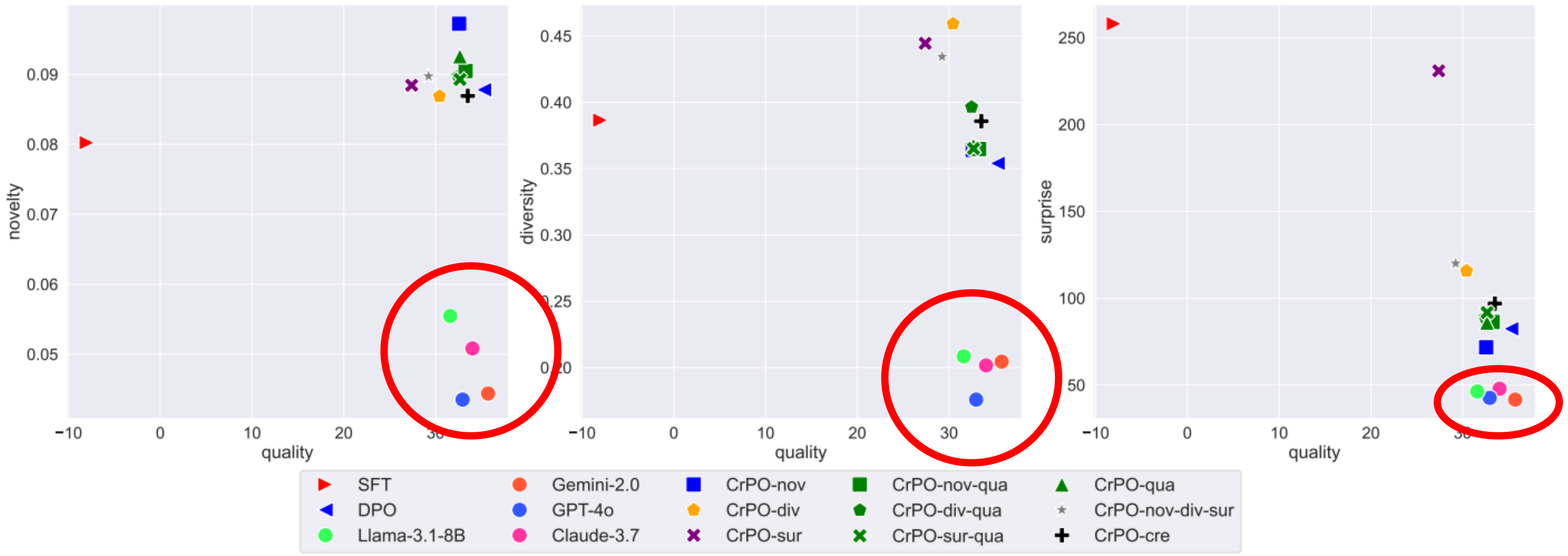


Figure 2: Results on held-out evaluation suite from MUCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3`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

Result

- the model trained with specific infection outperforms others

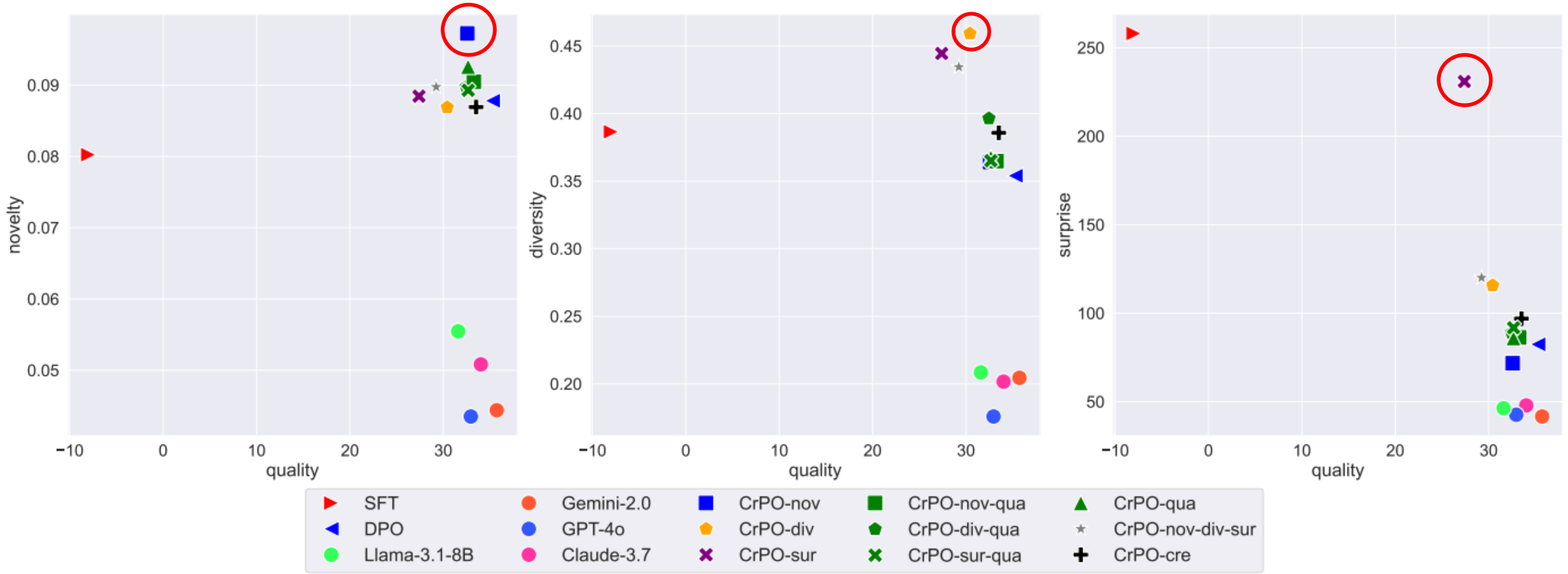


Figure 2: Results on held-out evaluation suite from MuCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3`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

Result

- Models that combine an external quality signal illustrates a trade-off

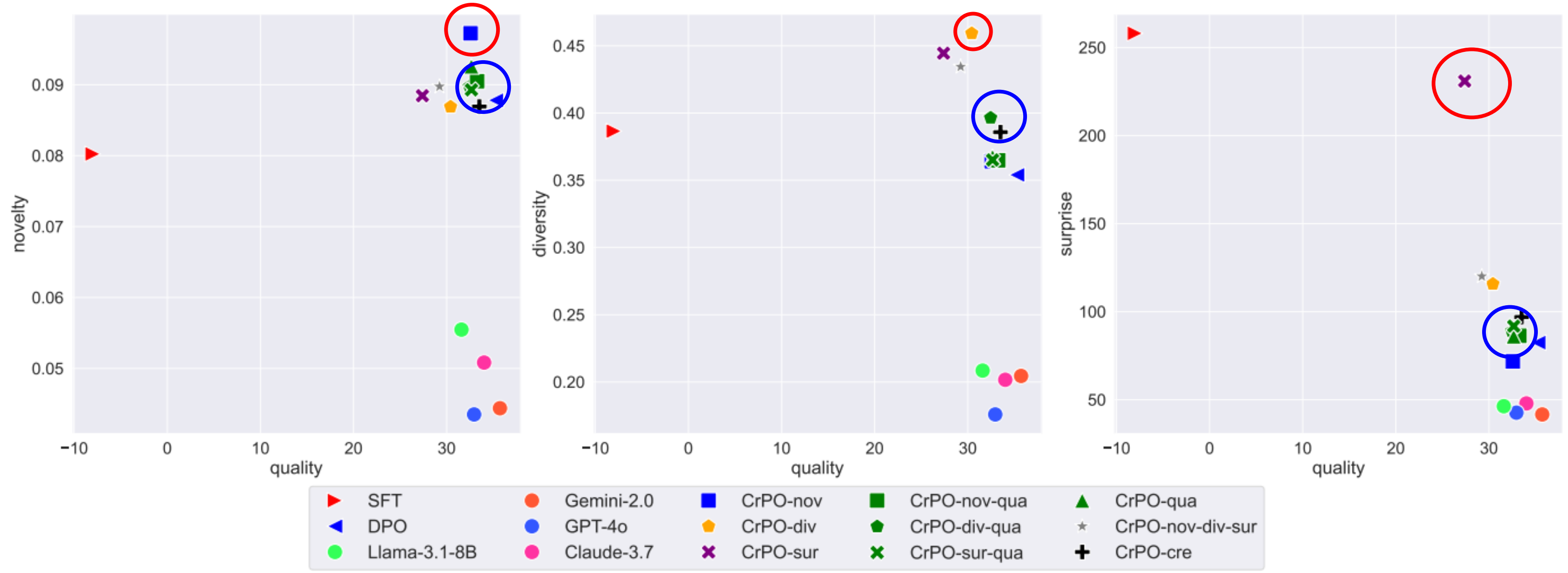


Figure 2: Results on held-out evaluation suite from MUCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3`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

Result

- Further highlights the balance between quality and other facets of creativity

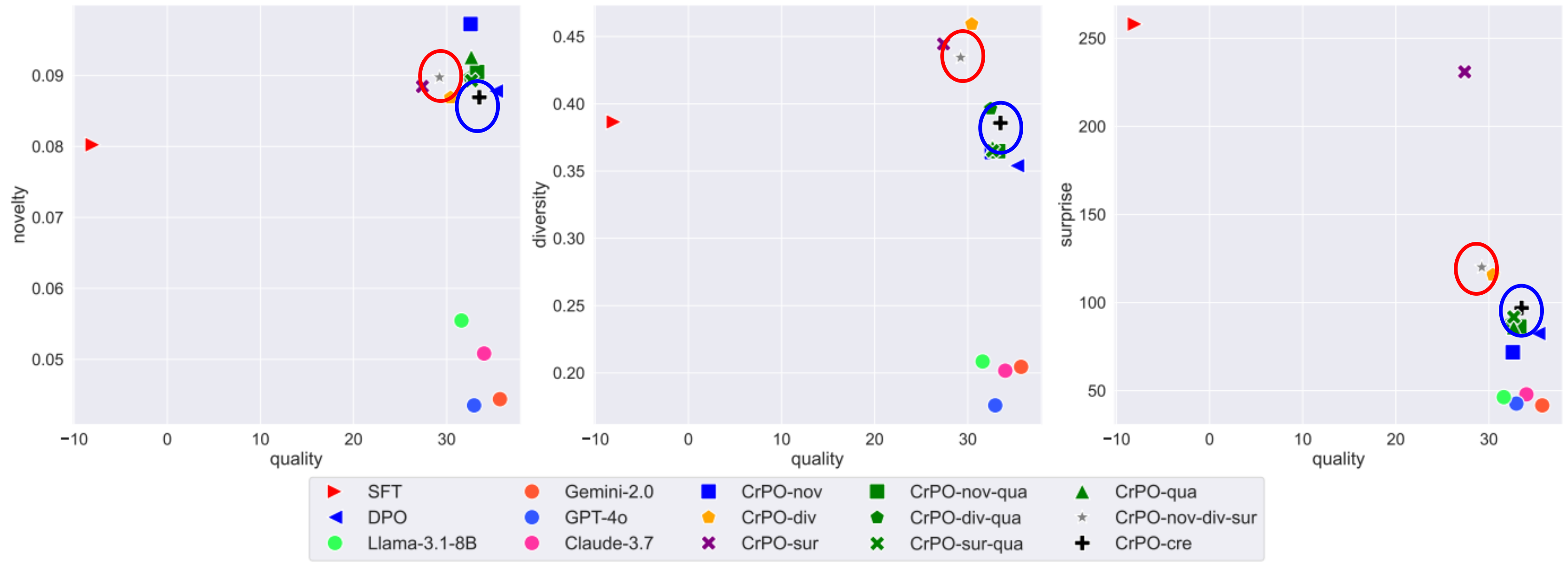


Figure 2: Results on held-out evaluation suite from MuCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3`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

Result

- Vanilla DPO model outperforms existing LLM baselines → strength of preference dataset

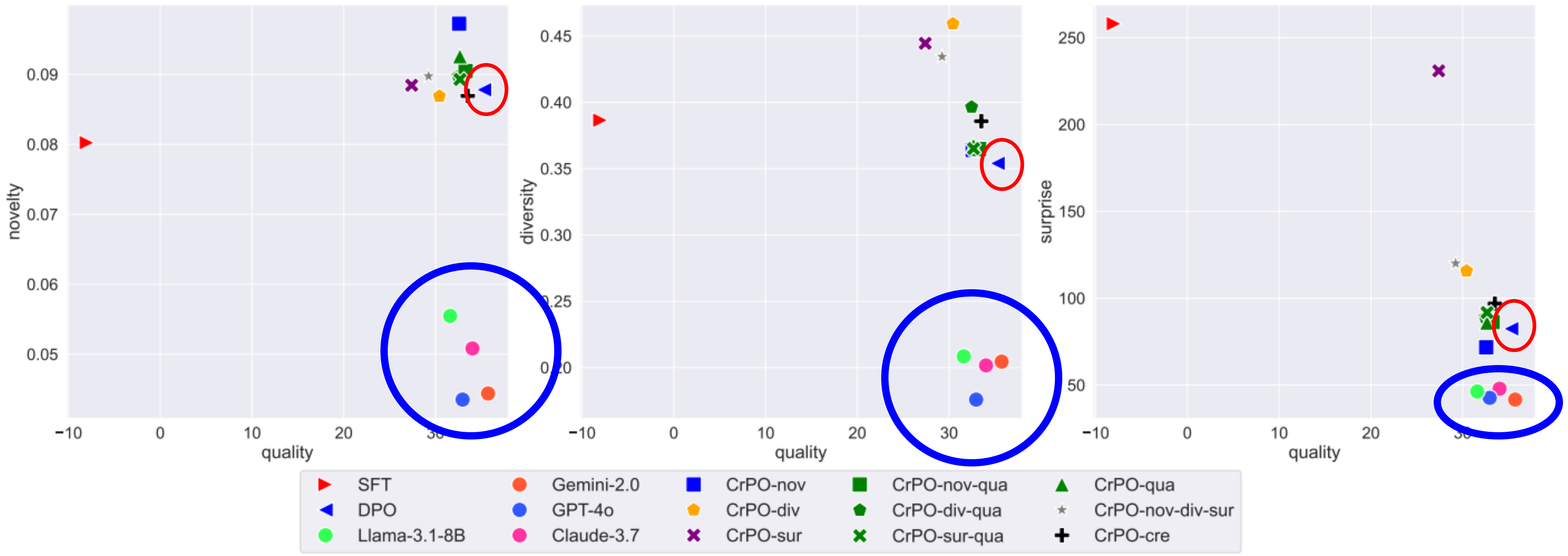


Figure 2: Results on held-out evaluation suite from MUCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3`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

Result

- SFT model performs worst in quality

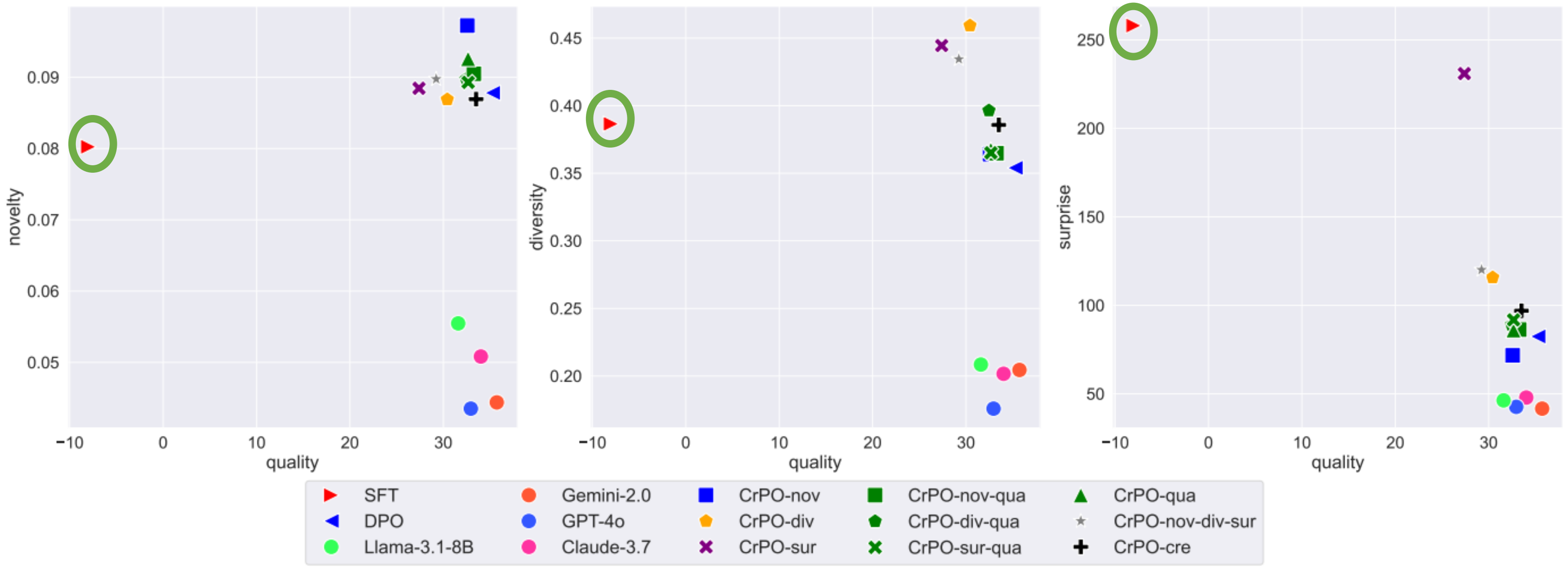


Figure 2: Results on held-out evaluation suite from MuCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3`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

Result: Effect of Infection Weights

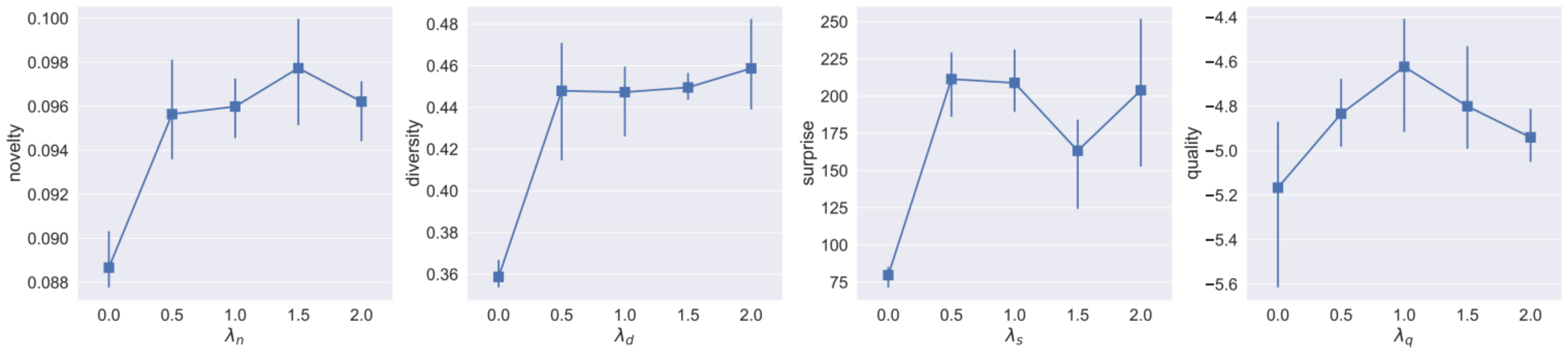
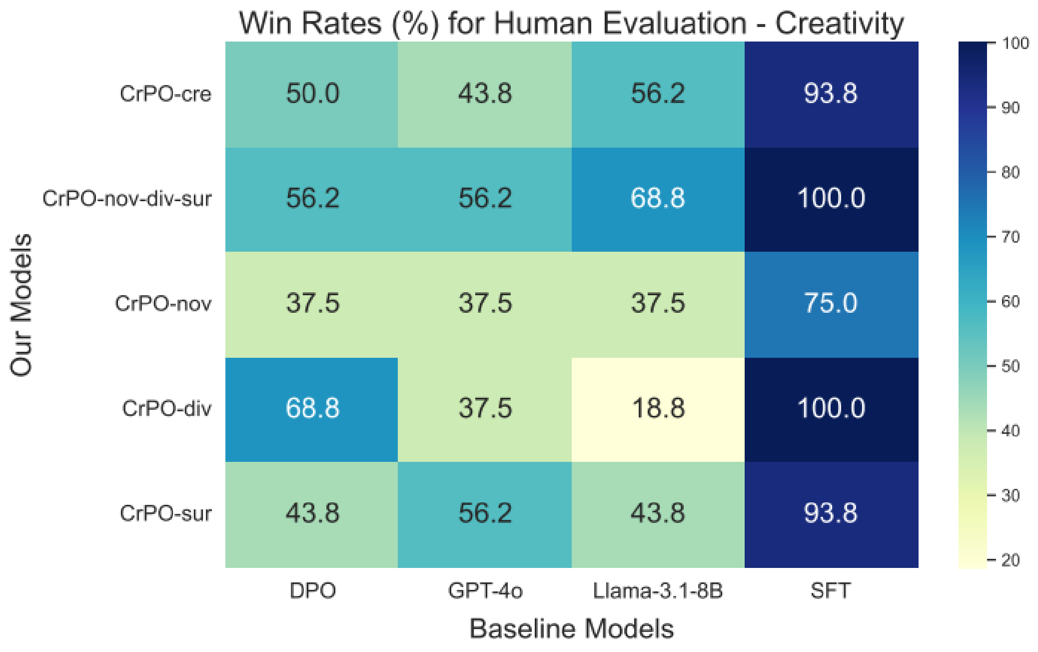


Figure 3: Effect of injection weights for each dimension. Results are averaged across three seed runs.

- an injection weight of 0.5 yields the greatest performance gains
- In terms of quality, the injection weight of 1.0 results in the highest performance

Result: Human Evaluation

- single task (Sentence Completion), 4 prompts, 4 baselines, and 5 CRPO variants
- CRPO-nov-div-sur model consistently outperforms all baselines
- In contrast, the full CRPO-cre model lags slightly



Result: Human Evaluation

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."

Result: NOVELTYBENCH Evaluation

- Includes tasks spanning randomness, factual knowledge, creative writing, and subjectivity
- clear separation: existing LLM baselines cluster VS our models
- the SFT model surprisingly achieves higher quality
  - 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

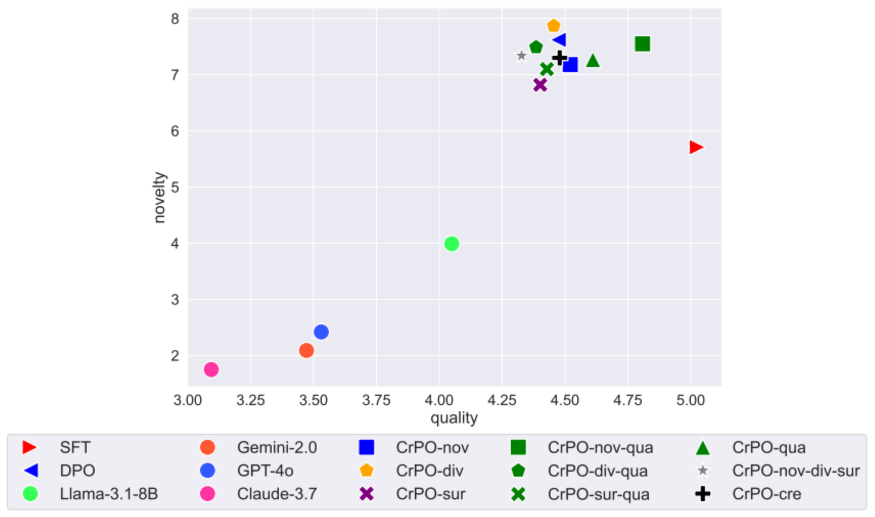


Figure 5: Evaluation results on NOVELTYBENCH, using the novelty and quality metrics defined in Zhang et al. (2025).

Conclusion

- CRPO Enhances LLM Creativity
  - CRPO: Method for aligning LLMs with human creative preferences
  - Improves novelty, diversity, surprise, and quality
  - Validated on MUCE & NOVELTYBENCH + confirmed by human raters
  - Outperforms strong baselines (e.g., GPT-4o, SFT, DPO)
  - Future Work: Scale to larger models, more languages, and other alignment methods

Additional – MUCE Dataset

- MUCE 데이터셋 어떻게 만들었는지
  - 수집 (Crowdsourcing + OSF)
    - 전 세계 창의성 연구자에게 직접 요청
    - Open Science Framework(OSF)에서 공개된 peer-reviewed 창의성 과업 데이터도 검색 relevant keywords (e.g., “creativity task”, “originality score”)

응답 ID	응답 내용	평가자 A 점수	평가자 B 점수	평가자 C 점수
R1	문을 고정하는 데 사용	4	4	5
R2	무기로 사용할 수 있음	2	3	2
R3	예술 작품을 만드는 재료	5	6	6

Additional – MUCE Dataset

- MUCE 데이터셋 어떻게 만들었는지
  - 수집 (Crowdsourcing + OSF)
    - 전 세계 창의성 연구자에게 직접 요청
    - Open Science Framework(OSF)에서 공개된 peer-reviewed 창의성 과업 데이터도 검색 relevant keywords (e.g., “creativity task”, “originality score”)

task	item	response	language	creativity_score
AlternativeUses	brick	문을 고정하는 데 사용	en	30
AlternativeUses	brick	예술 작품 재료	en	45

Additional – MUCE Dataset

- MUCE 데이터셋 어떻게 만들었는지

데이터셋	구성 요소	필터 기준
MUCE 원본	프롬프트 + 응답들 + 평가자별 점수	없음 (원천 데이터)
MUCE-PREF	프롬프트 + 선호 응답 + 비선호 응답	평가자 일치, 점수 차이 ≥ 5, 점수 ≥ 20, 쌍 수 ≤ 10
MUCE-SFT	프롬프트 + 선호 응답만 (output)	점수 ≥ 30인 preferred만 추출

Additional – MUCE Dataset

- Creativity가 어떤 과업들에서 어떤 형태로 나타났는지

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 2: 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
Consequences	5	1,315
Experiment Design	7	573
Hypothesis Generation	6	548
Research Questions	5	587
Associations	7	28
Total	**157**	**5,275**

Table 3: MUCE-SFT training dataset details.

Additional – MUCE Dataset

- Creativity가 어떤 과업들에서 어떤 형태로 나타났는지

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 10: MuCE dataset examples (Part 3).

Additional – NoveltyBENCH Dataset

Q. NoveltyBENCH 실험 결과를 넣은 이유?

A. 모델의 창의성 일반화 능력 평가

- MUCE 데이터셋에서 모델을 학습했지만, 동일한 분포의 데이터만 잘 다루는지, 아니면 창의성 과업 전반에 걸쳐 일반화가 가능한지를 확인하려면 외부 벤치마크 평가가 필요
- 따라서,최근 제안된 창의성 벤치마크인 NOVELTYBENCH를 사용하여 모델의 창의성 생성 능력이 다른 유형의 과업에서도 뛰어난지를 테스트하는 것이 주된 목적

Additional – NoveltyBENCH Dataset

Q. NoveltyBENCH 데이터셋은 어떤 데이터셋?

- NB-CURATED : Contains 100 prompts manually curated by the authors
- NB-WILDCHAT : Consists of 1,000 prompts automatically curated from real user interactions with ChatGPT

- 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?

Additional – NoveltyBENCH Dataset

Q. 두 평가양상이 왜 다른가? 특히 SFT는 quality 면에서 정반대의 결과

- 서로 다른 평가 지표 사용

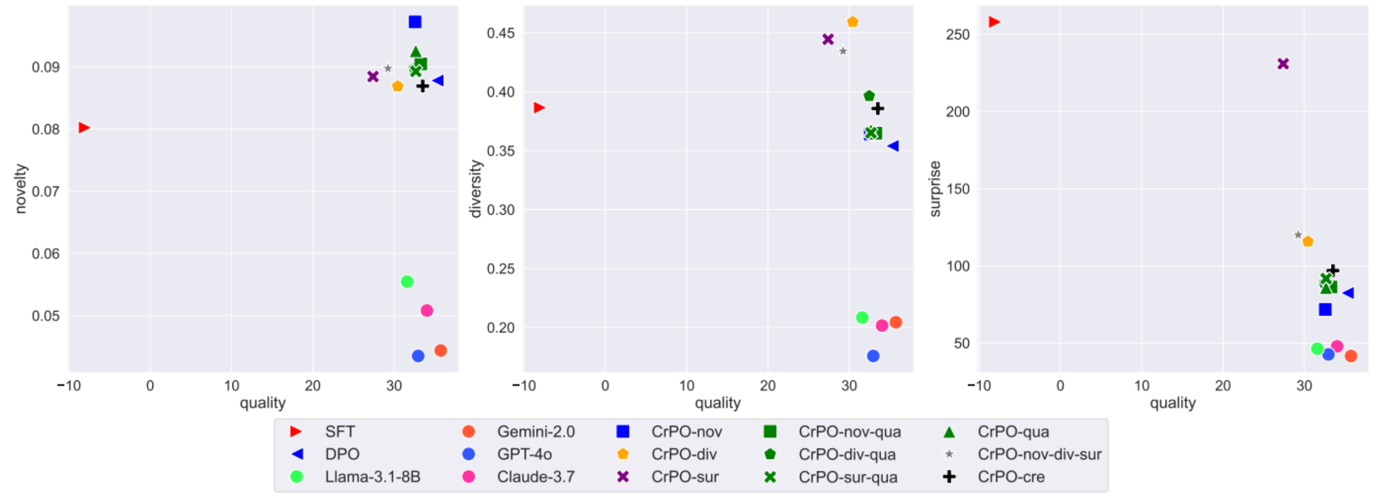


Figure 2: Results on held-out evaluation suite from MUCE across all baselines and our models using `Llama-3.1-8B-Instruct` as a base model. `nov`, `div`, `sur`, `qua`, `cre` denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. `Mistral-7B-Instruct-v0.3`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

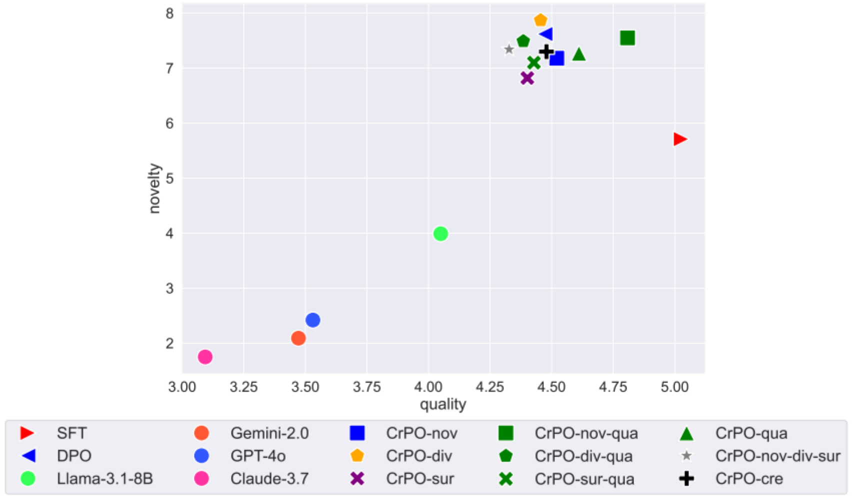


Figure 5: Evaluation results on NOVELTYBENCH, using the novelty and quality metrics defined in Zhang et al. (2025).

Additional – NoveltyBENCH Dataset

- NOVELTYBENCH의 평가지표: distinct, utility
  - Distinct: 주어진 프롬프트에 대해 모델이 k번 생성한 응답들 중, 기능적으로 서로 다른 응답 개수
  - 사람이 주석한 동등성 판별 데이터(1,100쌍)를 기반으로 훈련된 분류기(deberta-v3-large)가 각 응답 쌍이 "기능적으로 동일한지/다른지" 판단
  - Utility: 사용자가 모델로부터 최대 k개의 응답을 받는 상황을 가정하고, 각 응답의 기여도(효용)를 감안한 누적 점수

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

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

Additional – NoveltyBENCH Dataset

Q. 두 평가양상이 왜 다른가? 특히 SFT는 quality 면에서 정반대의 결과

- NoveltyBench의 quality는 “이 응답을 유저가 실제로 유용하다고 느낄까?”
- 즉, 형식이 명확하고 무난하며, 초반에 좋은 답을 보여주면 좋음
- 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

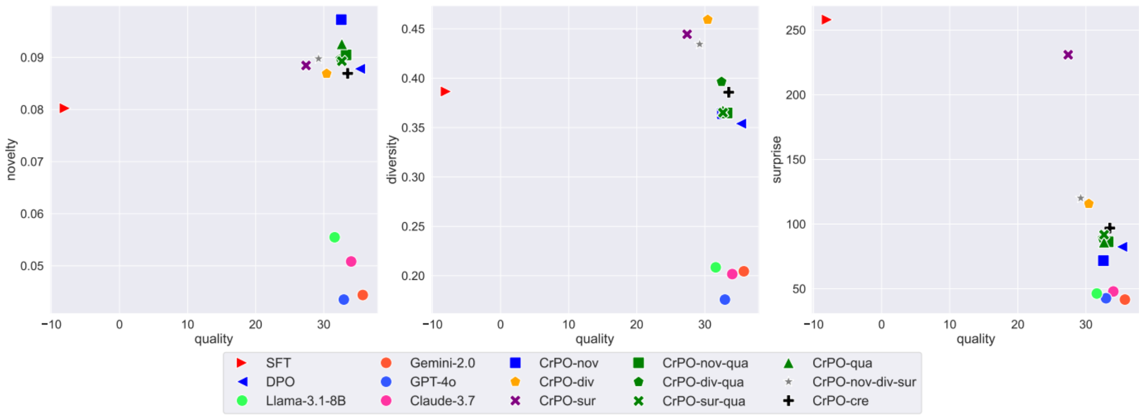


Figure 2: Results on held-out evaluation suite from MUCE across all baselines and our models using Llama-3.1-8B-Instruct as a base model. nov, div, sur, qua, cre denote novelty, diversity, surprise, quality, and creativity, respectively. Results are averaged across tasks. Mistral-7B-Instruct-v0.3 results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

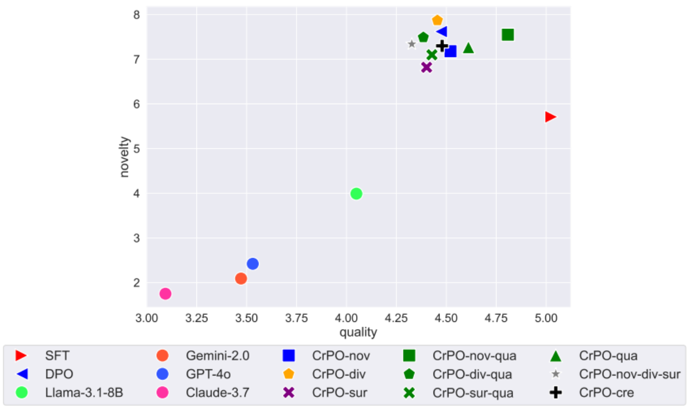


Figure 5: Evaluation results on NOVELTYBENCH, using the novelty and quality metrics defined in Zhang et al. (2025).

Thought?

- 기존의 문제 정의 → 방법 제안 → 실험 통한 효과성 입증까지 매끄럽게 이어져서 이해가 잘 됐다.
- 창의성 자체가 그렇긴 하지만, 정량 지표로는 뚜렷이 성능이 보이는데 정성 내용을 봐서는 차이가 뚜렷하진 않았다.
- 각각 surprise, novelty 등 지표 정의가 현 단계에서는 최선인 것 같지만 개선 여지가 있을 것 같음. 특히 novelty는 단어 쌍 간 참신성을 얼마나 서로 의미적 거리가 있냐를 보는데, quality랑 tradeoff가 큰 방식인 것 같음.
- DPO 복습이 돼서 좋았다

Open Question

- 어떻게 하면 단순한 의미적 거리를 넘어, 또한 품질과의 균형도 더 잘 유지할 수 있는 방식으로 novelty를 정의하고 평가할 수 있을까?