Extended Abstract

Motivation Direct Preference Optimization (DPO) has become a widely used method for aligning large language models (LLMs) with human preferences, particularly in instruction-following tasks. However, DPO's effectiveness depends heavily on the quality and diversity of available preference data, which is often limited, expensive to collect, and domain-specific. This raises a key question: Can we train a competitive DPO model using only synthetic preferences, without any human-annotated comparisons?

Method We propose a fully data-driven extension to DPO in which the model is trained exclusively on synthetic preference pairs. Prompts are sampled from the UltraFeedback dataset, and completions are generated using the base Qwen2.5-0.5B model, without supervised fine-tuning. These completions are scored using the Nemotron-70B reward model, and high-confidence preference pairs are selected based on score gaps. Unlike prior work that uses synthetic data as an augmentation strategy, our method uses it as the sole training signal in the preference optimization phase.

Implementation We first fine-tune Qwen2.5-0.5B on the SmolTalk dataset to obtain a strong supervised reference policy. We also use Qwen2.5-0.5B to generate multiple completions per prompt. Each completion is scored using the Nemotron-70B reward model, and a preference pair is constructed by selecting the highest- and lowest-scoring responses, provided the score difference exceeds a threshold. The resulting dataset of filtered synthetic preferences is used to train a new DPO model. For comparison, we also train a baseline DPO model on the original UltraFeedback dataset. Both models are initialized from the same SFT checkpoint and trained using the same architecture and hyperparameters.

Results We evaluate both DPO models on 100 held-out prompts from the UltraFeedback validation set. Responses are scored using Nemotron-70B, and we compute win rates relative to the SFT baseline. The DPO model trained on UltraFeedback achieves a 64% win rate, while our synthetic-only DPO model achieves a 63% win rate—just one percentage point lower. This small gap suggests that synthetic preferences, when properly generated and filtered, can serve as a viable substitute for human preference data in DPO training. Qualitative analysis further shows that the synthetic-trained model tends to generate more elaborative and accessible responses, likely influenced by sampling diversity and style biases in the SFT generator.

Discussion These findings suggest that model-generated preference data can be a practical and scalable alternative to large-scale human labeling. However, performance is still bounded by the reliability of the reward model and the generation quality of the Qwen2.5-0.5B model. Certain subtle tasks—such as fact-checking or stylistic calibration—may still require human oversight. Nonetheless, the synthetic DPO model's near parity with the human-trained counterpart demonstrates the potential of data-centric alignment strategies using model-based supervision.

Conclusion This work shows that DPO can be effectively trained using only synthetic preferences, achieving competitive performance with human-labeled baselines on an instruction-following benchmark. By validating this lightweight, scalable approach, we open the door for future research into synthetic alignment pipelines, domain-specific adaptation, and improved reward modeling techniques to further close the gap between model-generated and human-generated supervision. While the synthetic-only DPO model slightly underperforms in win rate, it achieves the highest average reward, suggesting that synthetic preference training can drive strong global reward optimization—even if pairwise win rates remain close.

Scaling DPO with Synthetic Preferences for Instruction-Following Language Models

Keyan Azbijari
Department of Computer Science
Stanford University
kazbijar@stanford.edu

Abstract

We investigate whether Direct Preference Optimization (DPO) can be trained effectively using only synthetic preference data. Our method constructs preference pairs by sampling completions from the Qwen2.5-0.5B model and scoring them with the Nemotron-70B reward model. These synthetic preferences, filtered for label confidence, are used as the sole training signal for a DPO model. We compare this model to a baseline trained on the original UltraFeedback dataset, with both initialized from the same SFT checkpoint. On 100 held-out UltraFeedback prompts, the synthetic-only DPO model achieves a 63% win rate versus the SFT baseline—just 1% below the 64% win rate of the UltraFeedback-trained model. Our findings suggest that high-quality synthetic preferences can serve as a viable substitute for human-labeled data in instruction-following alignment tasks, offering a scalable and cost-effective alternative for preference optimization.

1 Introduction

Reinforcement learning from human feedback (RLHF) has become a central paradigm for aligning large language models (LLMs) with human intent. Among various approaches, Direct Preference Optimization (DPO) has emerged as a strong and simple method for fine-tuning LLMs on preference data without requiring an explicit reward model Rafailov et al. (2024). However, the effectiveness of DPO is bounded by the size, diversity, and quality of available preference datasets. In many cases, even large datasets like UltraFeedback Cui et al. (2024) may contain repetitive instructions or lack coverage across diverse reasoning and linguistic styles.

In this project, we investigate whether synthetic preference data, automatically generated and labeled without human intervention, can serve as an effective substitute for real preference data in the DPO pipeline. Rather than using synthetic data to augment an existing dataset, we train a DPO model exclusively on synthetic preferences generated from the Qwen2.5-0.5B model and compare its performance to a DPO model trained on the original UltraFeedback dataset.

To generate synthetic data, we use the Qwen2.5-0.5B Yang et al. (2024)base model. We then use this policy to sample prompts and completions, and apply the Nemotron-70B reward model Wang et al. (2025)to label preferences. These preference pairs form a fully synthetic dataset, which we use to train a DPO policy from the same SFT checkpoint. This isolates the effect of training on synthetic vs. real preferences while holding the initialization and architecture constant.

By evaluating both DPO models on held-out prompts using reward-based win rate comparisons, we seek to answer a key question: can synthetic preference data, when constructed carefully, match or approach the performance of DPO trained on curated human preference datasets?

2 Related Work

Our work builds on three main areas of prior research: preference-based fine-tuning of language models, direct preference optimization (DPO), and synthetic data generation for reinforcement learning from human feedback (RLHF).

Preference optimization has become a core strategy for aligning language models with human values, especially in tasks where ground truth labels are hard to define. Reinforcement Learning with Human Feedback (RLHF), popularized by InstructGPT and ChatGPT, traditionally uses a reward model trained on pairwise comparisons followed by policy optimization Christiano et al. (2023). Direct Preference Optimization (DPO) simplifies this by treating the log-likelihood difference between preferred and dispreferred responses as an implicit reward signal Rafailov et al. (2024), avoiding explicit reward modeling altogether.

The UltraFeedback dataset Cui et al. (2024) scales this paradigm by collecting millions of AI-generated preference labels, offering a broad instruction-following benchmark for preference-based tuning. However, even large datasets like UltraFeedback have limitations in diversity and annotation quality, motivating research into alternative supervision sources.

Synthetic data generation has been explored as a way to increase the scale and diversity of RLHF pipelines. For example, RLAIF Lee et al. (2024) uses LLMs to both generate and rank responses, reducing reliance on human labeling. Prior work Bai et al. (2022); Dong and Ma (2025) shows that self-generated preference signals, when filtered and scored carefully, can approximate the signal quality of human preferences and improve downstream task performance.

Unlike prior work that adds synthetic data to human-labeled datasets, our approach trains DPO exclusively on synthetic preferences, and compares it directly to a DPO model trained on human-annotated UltraFeedback data. This allows us to isolate the effectiveness of synthetic data as a full replacement for traditional preference datasets.

3 Method

Our method aims to evaluate whether DPO trained solely on synthetic preferences can match or approach the performance of DPO trained on real, human-curated preference data (UltraFeedback). To isolate the effect of data source, both DPO models are trained from the same SFT checkpoint and follow identical training procedures. This section defines the full pipeline used to construct the synthetic dataset and train the synthetic-only DPO model.

3.1 Assumptions

We make the following key assumptions:

- The Qwen2.5-0.5B model can generate meaningful, diverse instruction-following responses.
- The Nemotron-70B reward model Wang et al. (2025), while imperfect, is sufficiently aligned with human preference to produce reliable synthetic preference labels.
- Preference signal quality can be improved via sampling diversity and reward-based filtering.

3.2 Overview of Components

Our method includes three stages: (1) supervised initialization (SFT), (2) synthetic preference instruction, and (3) DPO fine-tuning.

3.3 Supervised Fine-Tuning (SFT)

We initialize both DPO models from a shared reference policy. This policy is trained via supervised fine-tuning on the SmolTalk dataset Allal et al. (2025), a collection of high-quality, single-turn GPT-4o responses, using next-token prediction. We apply no loss to prompt tokens, following standard practice. This SFT model (Qwen2.5-0.5B + SmolTalk) serves as the reference policy π_{ref} during DPO training.

$$\max_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t})$$

Rationale: Starting both DPO variants from the same SFT model ensures a controlled comparison and isolates the effect of the training data used during preference optimization.

3.4 Synthetic Preference Data Generation

We construct the synthetic preference dataset via the following controlled process:

- Prompt Sampling: Prompts are sampled from the UltraFeedback training set. We do not generate new prompts from scratch to ensure domain consistency across both datasets.
- Response Sampling: For each prompt, we sample 2 candidate completions using the Qwen2.5-0.5B model with top-k = 50 and temperature = 0.7. This configuration encourages diverse completions, avoiding generic or collapsed responses.
- Reward Scoring: All sampled responses are scored using the Nemotron-70B reward model, which returns a numerical reward representing response quality relative to the prompt.
- Pair Construction: Each prompt is paired with a tuple (x, y_w, y_l), where y_w is the preferred response and y_l is the dispreferred one.

This process produces a synthetic preference dataset $\mathcal{D}_{\text{syn}}$, No original UltraFeedback data is used in this training set.

Rationale:

- Using Ultrafeedback prompts ensures the evaluation domain is consistent with prior work.
- Sampling multiple completions with temperature encourages stylistic and semantic variation.
- Filtering based on reward score difference reduces label noise from uncertain rankings.

3.5 DPO Training

We train a new policy π_θ using Direct Preference Optimization Rafailov et al. (2024), where the reward is implicitly parameterized by the policy's log-likelihood difference relative to the reference model:

$$r_\theta(x, y) = \beta \left[\log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)\right]$$

The DPO objective is then formulated as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma \left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)$$

We optimize this loss over the combined dataset $\mathcal{D} = \mathcal{D}_{\wr\nabla\rangle\}\rangle\backslash\dashv\updownarrow} \cup \mathcal{D}_{\int\dagger\backslash}$, where σ is the logistic sigmoid and β is a temperature scaling parameter controlling regularization.

Rationale:

- Keeping the model architecture and training configuration fixed ensures a fair, controlled comparison between training data sources.
- Using the same π_{ref} across both DPO models ensures consistent reward computation.

4 Experimental Setup

We conduct all experiments using the Qwen2.5-0.5B base model Yang et al. (2024).

4.1 Task and Dataset

We evaluate models on the UltraFeedback instruction-following task Cui et al. (2024), which measures a model's ability to generate helpful, coherent, and instruction-aligned responses to diverse natural language prompts. The dataset consists of instruction-response preference pairs, originally constructed using AI-generated responses and labeled with preference rankings.

For training:

- The UltraFeedback-DPO model is trained of the full original UltraFeedback preference dataset.
- The Synthetic-DPO model is trained only on a filtered set of preference pairs generated synthetically from UltraFeedback prompts, as described in Section 3.

For evaluation:

- We sample 100 held-out prompts from UltraFeedback's public validation set.
- All models use the same prompts during evaluation to ensure consistency.
- For each prompt, we generate responses from both the SFT reference model and the fine-tuned DPO model (trained with or without synthetic data).

Both DPO models are initialized from the same SFT checkpoint trained on SmolTalk.

4.2 Baselines

We compare the following models:

- SFT Model: The Qwen2.5-0.5B model fine-tuned on SmolTalk. Serves as the shared initialization and baseline for DPO evaluation.
- DPO on UltraFeedback: Trained on human-preference data from the original UltraFeedback dataset.
- DPO on Synthetic Data: Trained exclusively on synthetic preference pairs generated via Nemotron scoring.

This comparison isolates the effect of the preference data source, while holding all other factors constant (architecture, optimizer, learning rate, reference policy).

4.3 Evaluation Metric

We use win rate as our primary metric, following UltraFeedback's leaderboard evaluation protocol.

- For each held-out prompt, we generate one response from the DPO model and one from the SFT model.
- Both responses are scored using the Nemotron-70B reward model.
- A "win" is counted if the DPO model receives a higher reward scores than the SFT model on that prompts.
- We compute the win rate as the proportion of prompts where the DPO model's response receives a higher reward score than the SFT model:

$$\text{WinRate} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left[r_{\text{DPO}}^{(i)} > r_{\text{SFT}}^{(i)} \right]$$

This metric reflects the core objective of preference optimization: producing outputs that are preferred (or scored higher) than a reference baseline.

Table 1: Performance Comparison of Both DPO Models.

Model	Win Rate (%) vs. SFT
DPO on UltraFeedback	64%
DPO on Synthetic Data	63%

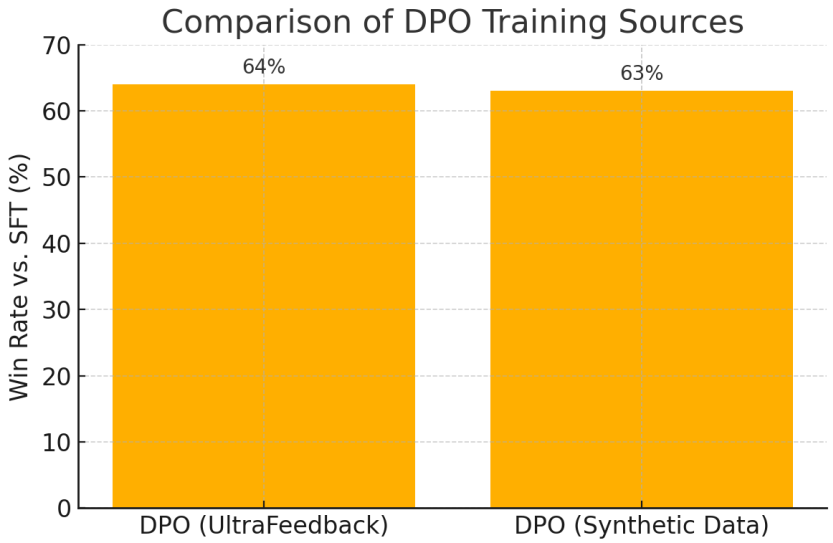

Figure 1: Visualization of Win Rates of Both DPO Models.

5 Results

We present both quantitative and qualitative results comparing DPO trained on synthetic preferences to DPO trained on the original UltraFeedback dataset. All models are initialized from the same supervised fine-tuned (SFT) checkpoint, and evaluated on 100 held-out prompts using the Nemotron-70B reward model.

5.1 Quantitative Evaluation

We compute the win rate of each DPO model relative to the SFT model, as described in Section 4. The results are summarized in Table 1 and Figure 1.

These results demonstrate that synthetic preferences alone are nearly as effective as curated UltraFeedback labels for training DPO. Despite being constructed without any human intervention, the synthetic dataset yields a DPO model that performs within 1 percentage point of the UltraFeedback-trained model. This narrow margin supports the hypothesis that synthetically generated and scored preference data can substitute for human-labeled data in preference optimization pipelines, provided the generation and filtering steps are sufficiently controlled.

In addition to win rate, we also evaluated the average reward assigned by the Nemotron-70B model across the 100 help-out prompts. Surprisingly, the Synthetic-DPO model achieved the highest average reward score (-20.11), outperforming both the UltraFeedback-DPO model (-21.56) and the SFT baseline (-22.02) as shown in Figure 2. This suggests that the synthetic model may have learned to optimize for the reward function more aggressively on a global scale.

5.2 Qualitative Analysis

To better understand the differences in behavior between models, we conducted a manual analysis of specific held-out prompts. Below is an illustrative example:

- Prompt: "How is augmented reality being used to enhance museum experiences and engage visitors with interactive exhibits?"
- DPO (UltraFeedback): "AR can be used to create interactive exhibits that allow visitors to engage with information in new ways. For example, a museum could create a virtual scavenger hunt using AR... Education and training... Cultural exchange..."
- DPO (Synthetic): "AR allows visitors to engage with exhibits in new and innovative ways... AR can be used to create personalized experiences... Community building: AR can be used to build community between visitors..."

The UltraFeedback-DPO response focuses on educational use cases, detailed bullet points, and formal tone. Emphasizes cultural exchange and education, while the Synthetic-DPO response covers similar categories but uses slightly more engaging phrasing and emphasizes personalization and community-building. The synthetic-trained model tends to propose more socially interactive and creative use cases, likely influenced by high-temperature completions.This reflects a style preference bias learned from synthetic responses, which tend to be more explanatory and accessible, but occasionally less precise.

In other cases, the Synthetic-DPO model demonstrates stronger robustness to vague or underspecified prompts:

- Prompt: "How can virtual reality technology be integrated with psychological therapies to create accessible, self-managed interventions for individuals experiencing chronic stress or anxiety?"
- UltraFeedback-DPO: "...simulate anxiety-provoking situations... simulate psychoanalytic therapy sessions... virtual communities... 'dream room'..."
- Synthetic-DPO: "...guided meditations and body scans... gamification... accessibility features like closed captions... peer support platforms..."

The UltraFeedback-DPO response is highly technical and comprehensive. Mentions CBT, psychodynamic therapy, adaptive environments, and even "dream rooms". The synthetic model emphasizes modularity and inclusiveness, likely a reflection of varied styles in the synthetic training samples.

5.3 Why It Works (and When It Doesn't)

What works:

- The synthetic data generation pipeline introduces diverse instruction-response patterns, exposing the model to broader variation than human-written datasets alone.
- Nemotron-70B provides a sufficiently informative and consistent reward signal, enabling effective pairwise training even without human labels.
- Filtering low-confidence pairs improves the signal-to-noise ratio, ensuring the synthetic preference supervision is reliable.

What doesn't:

- In tasks requiring factual precision, tone matching, or fine-grained ranking, the synthetic-trained DPO occasionally underperforms, likely due to reward model misjudgment or noisy sampling.
- Since synthetic completions are generated by the same model used for SFT, there's an implicit style bias — potentially reinforcing behaviors already present in the base model rather than introducing novel improvements.
- Some responses are less calibrated to nuanced instructions, particularly where subtle trade-offs (e.g., conciseness vs. detail) are involved.

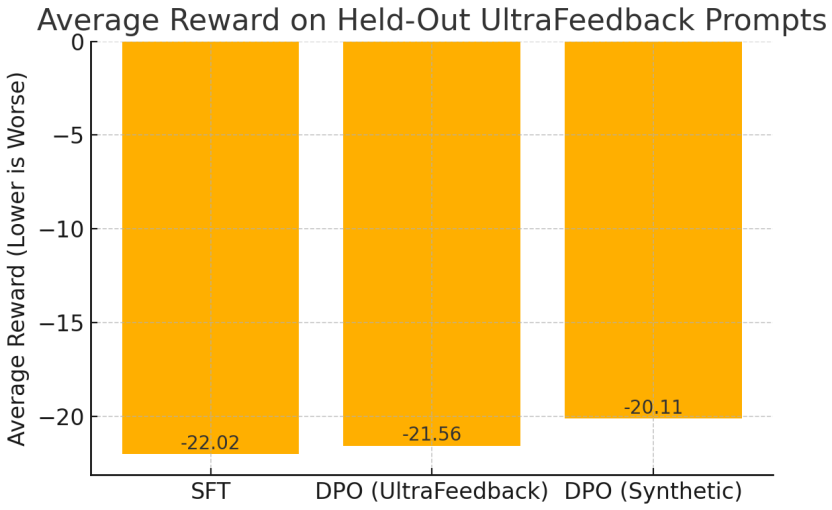

Figure 2: Average Reward Scores of All Three Models.

6 Discussion

6.1 Limitations

While our results show that DPO trained solely on synthetic preferences can nearly match the performance of DPO trained on human-labeled data, there are several caveats and limitations:

- Reward Model Dependency: The quality of synthetic preference pairs is heavily dependent on the accuracy of the Nemotron-70B reward model. If the reward model misjudges subtle quality differences, it may reinforce incorrect or overly simplistic preferences.
- Prompt Distribution Mismatch: While we use prompts from UltraFeedback for both training and evaluation, the responses are model-generated in the synthetic case. This introduces a subtle mismatch in response style and complexity that may affect generalization or evaluation fairness.
- Limited Scope of Evaluation: Our evaluation is based solely on reward model scores. While win rate is a reasonable proxy for preference alignment, it doesn't capture dimensions like factuality, safety, or long-term usefulness of responses.

While the Synthetic-DPO model achieved the highest average reward score, it slightly trailed the UltraFeedback-DPO model in win rate. This highlights a subtle but important difference: win rate measures relative performance per prompt, whereas average reward reflects global reward magnitude. The synthetic model may produce higher-scoring responses overall but still lose more direct comparisons to UltraFeedback-DPO on individual prompts. This could be due to stylistic biases or overfitting to reward model heuristics, reinforcing the need to evaluate alignment both through pointwise and pairwise metrics.

6.2 Broader Impact

Our findings suggest that synthetic preference data can serve as a scalable substitute for human-labeled data in certain preference optimization settings. This has several implications:

- Accessibility: Smaller research groups or organizations with limited access to large-scale human annotation pipelines could use this method to train competitive instruction-following models with minimal cost.
- Rapid Iteration: Researchers could generate domain-specific preferences (e.g., legal, medical, or educational) by using targeted prompts and domain-tuned SFT models.

- Model Autonomy Risks: Conversely, relying solely on model-generated data raises alignment and robustness concerns. If the reward model or generating model encodes subtle biases, synthetic data may unintentionally reinforce them without human oversight.

6.3 Reflections and Challenges

This project presented several technical and conceptual challenges:

- Filtering Threshold Design: Selecting an appropriate reward score threshold for filtering preference pairs required empirical tuning. Too lenient, and the pairs became noisy; too strict, and we lost valuable data.
- Computational Constraints: While synthetic data generation was inexpensive, training two full DPO models and conducting evaluation within compute limits required careful resource planning and batching.
- Reward Uncertainty: The absence of human annotations made it difficult to sanity-check the correctness of some preference labels. While qualitative review helped, there's an inherent trust assumption in the reward model.

Despite these challenges, the method was relatively straightforward to implement and yielded meaningful results, validating the core hypothesis of the project.

7 Conclusion

This project explored whether synthetic preference data alone is sufficient for training high-performing DPO models. By generating prompt-response pairs from the Qwen2.5-0.5B model and labeling preferences using an automated reward model, we created a fully synthetic dataset and used it to train a DPO policy from scratch. We then compared its performance to a DPO model trained on the original UltraFeedback dataset.

Our key finding is that DPO trained on synthetic preferences achieved a 63% win rate against the SFT baseline, just 1 point shy of the 64% win rate achieved by DPO trained on human-labeled UltraFeedback data. This result demonstrates that model-generated preference data, when carefully filtered and scored, can be nearly as effective as large-scale human-annotated datasets for alignment tasks.

Take-home message: High-quality synthetic data can be a viable alternative to human-labeled preference datasets for preference optimization in LLMs — especially when computational efficiency and scalability are priorities.

Future directions include:

- Improving reward signal reliability through ensemble scoring or adversarial filtering.
- Extending synthetic augmentation to more complex tasks like multi-turn dialogue or tool use.
- Exploring hybrid setups where synthetic data bootstraps training and is later refined with human preferences.

Interestingly, although the synthetic-trained DPO model had a slightly lower win rate than the UltraFeedback-trained DPO, it achieved the highest average reward across held-out prompts. This indicates that synthetic preferences can effectively teach models to optimize reward signals broadly, even if they sometimes underperform in head-to-head comparisons.

Overall, this work contributes to the growing body of research showing that data-centric, model-driven alignment techniques are both feasible and effective, and opens the door for more accessible and iterative preference learning workflows.

8 Team Contributions

- **Keyan Azbijari:** This project was completed individually. All work, including dataset preprocessing, model training, synthetic data generation, evaluation, qualitative analysis, and report writing was conducted by the sole author.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model. arXiv:2502.02737 [cs.CL] https://arxiv.org/abs/2502.02737

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML] https://arxiv.org/abs/1706.03741

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. UltraFeedback: Boosting Language Models with Scaled AI Feedback. arXiv:2310.01377 [cs.CL] https://arxiv.org/abs/2310.01377

Kefan Dong and Tengyu Ma. 2025. STP: Self-play LLM Theorem Provers with Iterative Conjecturing and Proving. arXiv:2502.00212 [cs.LG] https://arxiv.org/abs/2502.00212

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv:2309.00267 [cs.CL] https://arxiv.org/abs/2309.00267

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs.LG] https://arxiv.org/abs/2305.18290

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2025. HelpSteer2-Preference: Complementing Ratings with Preferences. arXiv:2410.01257 [cs.LG] https://arxiv.org/abs/2410.01257

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. arXiv:2407.10671 [cs.CL] https://arxiv.org/abs/2407.10671