The Paradox of Preference: A Study on LLM Alignment Algorithms and Data Acquisition Methods

Rishikesh Devanathan*, **Varun Nathan** and **Ayush Kumar**
{rishikesh.devanathan, varun.nathan, ayush}@observe.ai
Observe.AI
Bangalore, India

Abstract

This research investigates the impact of preference annotation acquisition methods on the performance of LLM alignment algorithms, including Direct Preference Optimization (DPO), Identity Preference Optimization (IPO), and Conservative DPO (cDPO), compared to Supervised Fine-Tuning (SFT) in NLP tasks. We analyze the influence of LLM and human-based preferences on algorithm performance, considering data volume and quality. Additionally, we assess DPO's vulnerability to overfitting and IPO's resilience against it, addressing four main research questions. Using the GAIR dataset and Zephyr-7b as the SFT model, we reveal unexpected negative outcomes. Specifically, DPO trained on LLM preferences outperforms human preferences, contrary to expectations. Moreover, there's no correlation between preference data volume or quality and algorithm performance. Contrary to expectations, DPO shows no overfitting in both human and LLM preference datasets. Surprisingly, cDPO doesn't fare better than DPO under flip noise. Our findings highlight the complexities of preference annotation methods and underscore the importance of scrutinizing negative results in NLP algorithm research.

1 Introduction

Large language models (LLMs) have proven their capacity to amass broad knowledge by simply maximizing the likelihood of human-written text but this objective isn't sufficient to generate responses that are safe, helpful and aligned with human preferences. Methods based on Reinforcement Learning with Human Feedback (RLHF), including Proximal Policy Optimization (PPO) (Schulman et al., 2017), aim to align LLMs with human preferences, a theme also explored in other papers (Ouyang et al., 2022; Askell et al., 2021; Bai et al., 2022a; Touvron et al., 2023). Direct Preference Optimization (DPO) (Rafailov et al., 2023) was later shown to train policies in a single stage, treating it as a classification task using human preference data. It's favored over PPO for its ability to handle reward translation issues well and consistently achieve high rewards across different levels of KL divergence in generated text.

Due to the expensive nature of collecting human annotations, LLM preferences serve as a substitute for human preferences in generating synthetic datasets (Chiang and Lee, 2023). Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022b) provides a promising alternative by leveraging a powerful off-the-shelf LLM to generate preferences for large-scale model training. The use of LLM preferences in dataset creation (Lee et al., 2023) has shown comparable performance between RLAIF and RLHF across various tasks, with performance degradation attributed to dataset quality issues, as evidenced by human-agreement scores. Conservative DPO (cDPO) [1] addresses these challenges by adopting a conservative target distribution, minimizing error probability, and deriving a loss function to ensure alignment between model preferences and observed preferences. The scarcity of diverse preference datasets poses a challenge for RLHF and feedback learning research. ULTRAFEEDBACK (Cui et al., 2023) addresses this challenge by providing an extensive, high-quality, and diversified preference dataset.

While widely adopted in preference optimization, DPO is susceptible to overfitting as observed by Tunstall et al. (2023) in the initial epoch of Zephyr-7B DPO training, but noted improved performance with further epochs. The IPO paper (Azar et al., 2023) discovered that RLHF and DPO are prone to overfitting due to relying on the assumption that pairwise preferences can replace ELo-

* Work done during internship at Observe.AI

[1] https://ericmitchell.ai/cdpo.pdf

Proceedings of the Fifth Workshop on Insights from Negative Results in NLP, pages 135–147

scores through Bradley-Terry modeling. To mitigate this, IPO introduces a regularizing term controlling log-likelihood ratios to address overfitting to the preference dataset.

A literature gap exists in exploring how training data volume influences LLM alignment algorithms, DPO and IPO. Empirical evidence is lacking on IPO's ability to counter DPO's overfitting, and studies on data quality's impact on DPO, and cDPO's effectiveness in addressing it, are scarce. It's essential to investigate the influence of preference annotation methods (LLM vs. human preferences) on these factors and the performance of LLM alignment algorithms, including DPO, cDPO, and IPO, given the increasing use of LLM preferences.

In this work, we investigate how the method of preference annotation acquisition affects the critical performance factors influencing the effectiveness of LLM alignment algorithms and seek to address the following research questions:

- **RQ1:** How does the choice of preference annotation acquisition method influence the performance of DPO and IPO in comparison to SFT?
- **RQ2:** What is the effect of data volume on the performance of DPO and IPO? Does the relationship depend on the preference annotation acquisition method?
- **RQ3:** What is the effect of data quality on the performance of DPO and cDPO? Does the relationship depend on the preference annotation acquisition method?
- **RQ4:** To what extent does DPO suffer from overfitting, and can IPO withstand it? How does the preference annotation acquisition method impact this phenomenon concerning both loss functions?

We demonstrate unexpected superiority of LLM trained with DPO on LLM preferences over human preferences. Performance shows no correlation with data volume or quality. DPO doesn't exhibit overfitting issues, while cDPO doesn't improve under noise. Our findings highlight challenges in preference annotation and aligning LLMs.

2 Implementation Details

We choose Zephyr-7B as our SFT model and GAIR (Li et al., 2024) as our preference dataset, containing both human and LLM preferences, for our experiments. MT Bench (Zheng et al., 2023) is used to evaluate our models while GPT4-Turbo is chosen as the LLM for obtaining synthetic preferences. Further details on our choices are provided in Section A.1, including hyperparameter specifics.

3 Results and Analysis

In this section, we provide a comprehensive analysis of the performance evaluation results, shedding light on the key observations made during our study.

3.1 RQ1 (Preference model performance)

To investigate this, we independently fine-tuned Zephyr-7B (SFT) using preferences from both GPT4-Turbo and humans in the GAIR dataset. In Table 1, the IPO model trained on human preferences, as anticipated, outperforms its GPT4-Turbo-trained counterpart according to the MT Bench score. However, contrary to expectations, the DPO model trained on GPT4-Turbo preferences outperforms its human-trained counterpart according to the MT Bench score. We speculate that GPT4, acting as the MT Bench judge, might show bias towards responses from GPT4-Turbo-trained models. To verify, we collect predictions from both models on MT Bench, comprising 160 samples, and shared them with our in-house annotation team of three members. Model names were concealed, and annotators chose from options 'model 1' (GPT4-Turbo preference trained model), 'model 2' (human preference trained model), or 'equal' based on the quality of the generated output. We opt for a majority vote to determine the final preference, and the inter-annotator agreement score, calculated using Fleiss' Kappa (Fleiss et al., 1971), was measured at 0.64. As shown in Table 2, the model trained on GPT4-Turbo preferences was preferred in 63 of 160 samples with a much higher win rate of 39.4%, suggesting alignment between MT Bench scores and human annotation. This negative outcome of model trained on LLM preference data outperforming the one trained on human preference data prompts a crucial inquiry regarding the superiority of human preferences over those sourced from LLMs and the necessary measures to guarantee the quality standards of human-collected data.

3.2 RQ2 (Effect of data volume)

To investigate this, we independently fine-tune Zephyr-7B (SFT) using DPO and IPO losses on

Algorithm	GPT4-Turbo	Human
SFT (Baseline)	6.753	6.753
Preference Model (DPO)	6.994	6.722
Preference Model (IPO)	5.125	5.484

Table 1: Benchmarking performance of DPO and IPO by Preference annotation acquisition method using MT Bench scores

Model	# Wins	# Ties	Win Rate
DPO (GPT4-Turbo)	63	51	39.38%
DPO (Human)	46	51	28.75%

Table 2: Results from human annotation of DPO model trained on GPT4-Turbo and Human preferences

Data Volume (% Train Data)	GPT4-Turbo		Human	
	Loss = DPO	Loss = IPO	Loss = DPO	Loss = IPO
100%	6.994	5.125	6.722	5.484
75%	6.756	5.741	6.544	6.300
50%	6.878	6.766	6.897	6.692
25%	6.788	6.953	6.819	6.928

Table 3: Benchmarking performance of DPO and IPO models by preference annotation acquisition method when trained on different data volumes using MT Bench scores

sampled datasets with varying proportions of preferences from both GPT4-Turbo and humans in the GAIR dataset. Contrary to the anticipation of improved generalization with increased data diversity, this pattern is absent in DPO and IPO models trained on both types of preferences (Table 3). Neither GPT4-Turbo-Preference-trained nor human-preference-trained DPO and IPO models demonstrate a monotonic relationship with data volume, suggesting that augmenting preference data volume may not necessarily enhance model performance. Notably, DPO and IPO models trained on 25% human preference data outperform those trained on the entire dataset, hinting at potential overfitting issues. We conduct an exhaustive examination into the susceptibility of DPO to overfitting, with detailed results emphasized in 3.4.

Table 3 also demonstrates that models trained with IPO underperform those trained with DPO across sample proportions of 100%, 75%, and 50% in both GPT4-Turbo and human preference datasets. Upon conducting a hyperparameter sweep over a fine-grained range for the DPO and IPO models trained on 100% of the human preference dataset, significant uplift in performance was observed for both IPO and DPO models post-tuning β as indicated in Table 8. However, we see that DPO still outperforms IPO, indicating the inefficacy of IPO in surpassing DPO despite extensive tuning of β. Due to the high cost of running these experiments and the limited effectiveness of IPO, we did not extend the tuning exercise to other configurations.

This finding highlights the valuable insight for ML researchers and scientists in enterprises using DPO for preference modeling. It also underscores the challenges involved in exploring alternative loss functions such as IPO to enhance performance with limited preference data.

3.3 RQ3 (Effect of data quality)

To tackle this issue, we independently fine-tune Zephyr-7B (SFT) using DPO and cDPO losses on sampled datasets with varying levels of flip noise introduced into preferences from both GPT4-Turbo and humans in the GAIR dataset. Flip noise is introduced by swapping the chosen response and the rejected response for a selected percentage of prompts. Despite the anticipation that models would exhibit better generalization with higher data quality, this pattern is not evident in DPO and cDPO models trained on both types of preferences (Table 4). Intriguingly, DPO models trained with 25% flip noise outperform those trained on clean data across GPT4-Turbo and human preferences, while the cDPO model only marginally outperforms it when trained on 50% flip noise data.

Moreover, Table 4 indicates that models trained with cDPO consistently exhibit inferior performance compared to those trained with DPO across all configurations and datasets. This contradicts expectations set by the cDPO paper, which suggests that cDPO's ability to optimize to a fixed delta from the reference model and then halt likely enhances its stability compared to the original DPO loss, making it more effective when dealing with noisy

Data Quality (% Flip Noise)	GPT4-Turbo		Human	
	Loss = DPO	Loss = cDPO	Loss = DPO	Loss = cDPO
0%	6.994	6.994	6.722	6.722
5%	6.956	6.733	6.759	6.559
25%	7.013	6.313	6.731	6.284
50%	7.081	6.372	6.703	6.344
75%	6.984	5.456	6.584	5.378

Table 4: Benchmarking performance of DPO and cDPO models by preference annotation acquisition method when trained on datasets with different flip noise ratios using MT Bench scores

# Steps	GPT4-Turbo				Human							
	Loss = DPO		Loss = IPO		Loss = DPO		Loss = DPO (Tuned)		Loss = IPO		Loss = IPO (Tuned)	
	Training Loss	MT Bench Score	Training Loss	MT Bench Score	Training Loss	MT Bench Score	Training Loss	MT Bench Score	Training Loss	MT Bench Score	Training Loss	MT Bench Score
1/4	0.212	6.809	11.796	6.578	0.283	6.638	0.539	6.969	14.775	6.669	0.577	6.919
2/4	0.037	6.859	5.959	5.244	0.057	6.744	0.505	6.919	9.322	4.677	0.368	7.056
3/4	0.024	6.813	4.056	4.781	0.038	6.728	0.253	6.981	6.648	5.874	0.274	6.953
4/4	0.018	6.994	3.231	5.125	0.031	6.722	0.219	7.184	5.415	5.484	0.241	7.113

Table 5: Impact of Overfitting on DPO and IPO Models at 100% Data Volume by Preference Annotation Acquisition Method

training data. Upon conducting a thorough hyperparameter sweep over a finely grained range for both DPO and cDPO models trained on the human preference dataset with 5% flip noise, significant performance enhancements were observed for both after β tuning as indicated in Table 9. However, DPO continues to surpass cDPO, indicating the limited efficacy of cDPO even after extensive β tuning. Due to the substantial expenses involved in running these experiments and the limited effectiveness of cDPO, we discontinued extending the tuning process to other configurations.

This negative outcome holds considerable significance for researchers and professionals in organizations utilizing DPO for preference modeling in noisy datasets.

3.4 RQ4 (DPO and IPO overfitting)

Our objective is to validate the hypothesis that DPO is susceptibile to overfitting and IPO is resilient against it (Azar et al., 2023). We conduct empirical validation by independently fine-tuning Zephyr-7B (SFT) using DPO and IPO losses on 100% of preferences from both GPT4-Turbo and humans in the GAIR dataset. Overfitting is assessed by monitoring training loss and evaluation scores of checkpoints at intervals of 25% of the training steps on MT Bench.

Table 5 reveals that DPO exhibits overfitting only when trained on human preferences with the default β of 0.1, contrasting IPO, which exhibits overfitting when trained on both types of preferences. As suggested in the paper, we hypothesised that tuning beta would help mitigate overfitting in IPO trained model. As expected, when examining models trained with a tuned β, a different pattern emerges, where both DPO and IPO models trained on human preference data do not display overfitting. Thus, the key negative result we observe is that tuning β (0.00625) helps mitigate overfitting in DPO when trained on human preferences, providing valuable insights for ML researchers and industry practitioners employing DPO and IPO for preference modeling with limited data.

4 Conclusion

We analyze the influence of data quantity on DPO and IPO, utilizing LLM preferences and human preferences. Surprisingly, there's no linear correlation between data quantity and performance. Similarly, the impact of data quality on DPO and cDPO, using both LLM preferences and human preferences, also lacks a linear trend with performance. Contrary to expectations, DPO trained on LLM preferences outperforms its human-trained counterpart. Additionally, IPO fails to outperform DPO across various data volumes, while cDPO struggles to address induced flip noise in preferences. Interestingly, DPO shows no signs of overfitting when trained on both LLM and human preference datasets. These findings prompt further research to enhance the resilience and effectiveness of LLM alignment algorithms in preference modeling.

5 Limitations

This study offers significant insights into the performance of LLM alignment algorithms and the influence of preference annotation acquisition methods, but it is not without its limitations. First, the research is grounded in a specific set of LLM alignment algorithms, namely DPO, IPO, and cDPO. The results may not extend to other alignment algorithms like KTO (Ethayarajh et al., 2024). Future studies could broaden the scope by examining the performance of different algorithms for a more holistic understanding of the field. Second, the GAIR training dataset and MT Bench evaluation dataset were used in this study. The outcomes might vary with the use of different datasets, hence, extrapolating these findings to other contexts should be done with caution. Third, the Zephyr-7b, a decoder-only model, was used as the underlying SFT model, and GPT4-Turbo was used as the source in GAIR for acquiring LLM-based preferences. The outcomes might differ with the use of other models. Specifically, the trends observed may not necessarily apply to other SFT models within the same architectural class or different architectural classes such as encoder-decoder models. Fourth, the study did not find a correlation between the volume or quality of preference data and algorithm performance. However, this does not exclude the possibility of other factors influencing algorithm performance. Additional research is required to identify these potential factors. Fifth, the study found that DPO trained on LLM preferences outperforms human preferences, which was unexpected. This raises questions about the validity of human preferences as a performance benchmark for algorithms. Future research should delve deeper into this issue. Lastly, the study found no evidence of overfitting in DPO when trained on both LLM and human preference datasets. However, this finding should be interpreted cautiously as overfitting is a multifaceted issue influenced by various factors, including model complexity, training dataset size, and data noise. Further research is needed to fully comprehend the conditions that may lead to overfitting.

In conclusion, while this study offers valuable insights into the performance of LLM alignment algorithms and the impact of preference annotation acquisition methods, these findings should be considered in light of the aforementioned limitations. Future research should strive to address these limitations for a more comprehensive understanding of the field.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *CoRR*, abs/2310.12036.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

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 Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí 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. 2022b. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15607–15631. Association for Computational Linguistics.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and

Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *CoRR*, abs/2310.01377.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with *V*-usable information. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: model alignment as prospect theoretic optimization. *CoRR*, abs/2402.01306.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. RLAIF: scaling reinforcement learning from human feedback with AI feedback. *CoRR*, abs/2309.00267.

Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. Dissecting human and LLM preferences. *CoRR*, abs/2402.11296.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *CoRR*, abs/2310.16944.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Appendix

A.1 Implementation Details for LLM Alignment Experiments

In this section, we elaborate on the implementation details of our study, exploring how variations in the quality and quantity of preference data impact the performance of DPO, IPO, and cDPO, alongside the influence of preference annotation acquisition methods.

We opt for Zephyr-7B, a decoder-only model based on Mistral-7B, as our SFT model due to its top-ranking performance in MT Bench (top 5 in the list of models with a non-proprietary license) and accessibility in the HuggingFace model repository under the apache-2.0 license.

We employ the GAIR preference dataset (Li et al., 2024), comprising 5.24K curated conver-

sations with pairwise human preferences from 13K unique IP addresses on the Chatbot Arena, collected between April and June 2023. Additionally, binary preference labels are gathered from 32 LLMs, incorporating 2 proprietary and 30 open-source models. With 29 defined properties, each response is annotated using Likert scale ratings or property-specific annotations. This dataset is selected primarily for its inclusion of both human and LLM preferences. Furthermore, the renowned LIMA paper (Zhou et al., 2023) originates from the same organization that released this dataset.

We also conduct experiments on two additional preference datasets: Ultrafeedback (Cui et al., 2023), comprising 61K prompts with preferences sourced from GPT4, and the Stanford Human Preferences Dataset (SHP) (Ethayarajh et al., 2022), extracted from posts and user comments across 18 subreddits containing human preferences, totaling 349K samples, which we downsample to 100K samples by filtering for those with a score ratio greater than 2 for experimentation. These datasets are selected for their extensive scale and diversity compared to other datasets.

Models are evaluated using MT Bench (Zheng et al., 2023), a curated benchmark featuring 80 high-quality, multi-turn questions designed to evaluate conversation flow and instruction-following capabilities in multi-turn dialogues. GPT-4 rates MT Bench outputs on a scale of 1-10, with higher scores indicating better performance. Refer to Table 15 for the domains considered in the datasets. Average MT Bench scores across questions and turns are reported for all experiments.

We fine-tune all alignment models for two to three epochs, following the approach in Tunstall et al. (2023). Adam optimizer with betas of (0.9, 0.999) and epsilon of 1e-08 is utilized. A linear learning rate scheduler with a peak rate of 5e-7 and 10% warmup steps is applied. Models are trained with a global batch size of 16, using β = 0.1 to control deviation from the reference model. A hyper-parameter sweep for β is performed over the range $\in \{1e-3, 2.5e-3, 5e-3, 6.25e-3, 1e-2, 5e-2, 1e-1, 1.5e-1, 2e-1, 5e-1, 9e-1\}$ for four settings: training DPO / IPO models on 100% data volume + 0% flip noise and DPO / cDPO models on 100% data volume + 5% flip noise. β tuning is specifically focused on due to its significant impact on model performance. Given the high training cost, β tuning is not conducted for all experiments. Experiments are conducted on an AWS p4de.24xlarge instance with eight GPUs, each with 80 GB of memory. A single training run takes 3-4 hours on average, costing approximately $140-190. Results are reported as the mean of 4 runs.

Dataset: https://huggingface.co/datasets/GAIR/preference-dissection

Training Code: https://github.com/huggingface/alignment-handbook/tree/main

Evaluation Code: https://github.com/lm-sys/FastChat

A.2 Error Analysis

In our study, we encountered several unexpected results that contradicted our initial hypotheses. This section provides an in-depth error analysis to understand these observations and their potential causes.

Firstly, we posited that DPO performance would be superior when trained on human preferences compared to LLM preferences. However, our findings contradicted this hypothesis. One plausible explanation for this unexpected outcome could be the inherent biases present in human preferences, which may not align with the objective function of the DPO algorithm. Moreover, there may be inherent limitations in the methodology used to collect human annotations.

An example of this discrepancy is evident in the performance of the DPO model trained on GPT4-Turbo preferences versus human preferences, particularly in the task of coding, as illustrated in figures 1 and 2. It is conceivable that the expertise levels of the human annotators selected for this task were not carefully considered.

Additionally, our analysis revealed instances of hallucinations (Row 3 in Table 16) and the generation of incomplete or redundant responses for Math questions (Row 6 in Table 16) by the DPO model trained on human preferences. These discrepancies may be attributed to various biases inherent in human preferences or inconsistencies in annotation practices.

Conversely, LLM preferences may exhibit greater consistency or comprehensiveness, thereby yielding superior performance. Further investigation is warranted to elucidate this discrepancy.

Secondly, we observed no discernible correlation between the volume or quality of preference data and the performance of the alignment algorithms. This finding challenges the widely held assumption that larger, higher-quality datasets in-

variably lead to improved performance. One potential contributing factor to this discrepancy could be the possible absence of independence and identical distribution (iid) in the data sourced from GAIR (Li et al., 2024), which may have influenced the outcomes of our experiments.

As depicted in figures 3 and 4, the disparities in performance among models trained on varying data volumes or with different proportions of flip noise are not uniformly distributed across the domains in MT Bench (Zheng et al., 2023). To delve deeper into this phenomenon, we manually mapped the domains in GAIR (Li et al., 2024) and MT Bench (Zheng et al., 2023), as illustrated in Table 6. Subsequently, we aggregated the data volume from GAIR (Li et al., 2024) according to the distinct domains in MT Bench (Zheng et al., 2023), as presented in Table 7.

As demonstrated in Table 7, there exists an imbalance in the distribution of samples within GAIR (Li et al., 2024) across the various domains in MT Bench (Zheng et al., 2023). This non-uniform distribution could potentially skew the results of our experiments on data quantity and quality. Additionally, it is plausible that there are diminishing returns once a certain threshold of data volume is surpassed.

Thirdly, contrary to our initial expectations, the DPO algorithm did not exhibit indications of overfitting on either the human or LLM preference datasets. This suggests the possibility that our methods for detecting overfitting may not have been sufficiently sensitive. Moreover, the relatively small volume of the GAIR (Li et al., 2024) dataset, consisting of approximately 5.2K samples, may have biased the results pertaining to DPO overfitting. It is conceivable that the dataset lacked the requisite data volume to effectively capture the onset of overfitting.

Furthermore, the decision to train for only 2-3 epochs might have been too brief to provoke overfitting, particularly because the learning rate was appropriately set. We opted for this duration based on the findings of Tunstall et al. (2023), who reported observing overfitting after a single epoch. However, the divergence in observed behaviors could be attributed to the differences in the nature and size of the datasets.

It is worth noting that models often necessitate additional training iterations before overfitting manifests, as they gradually adapt not only to the underlying pattern but also to the noise present in the training data. Consequently, further investigation is warranted to ascertain the precise underlying cause of these observations.

Lastly, cDPO did not perform better than DPO under flip noise conditions. This was surprising as cDPO is designed to be more conservative and thus more resilient to noise. One possible explanation could be that the flip noise in our dataset was not significant enough to differentiate the performance of DPO and cDPO. Alternatively, there might be other types of noise or errors that cDPO is not equipped to handle.

In conclusion, our error analysis has revealed several unexpected findings that challenge common assumptions in LLM alignment algorithm research. These findings underscore the importance of rigorous error analysis and the need for further research to understand the complexities of preference annotation methods.

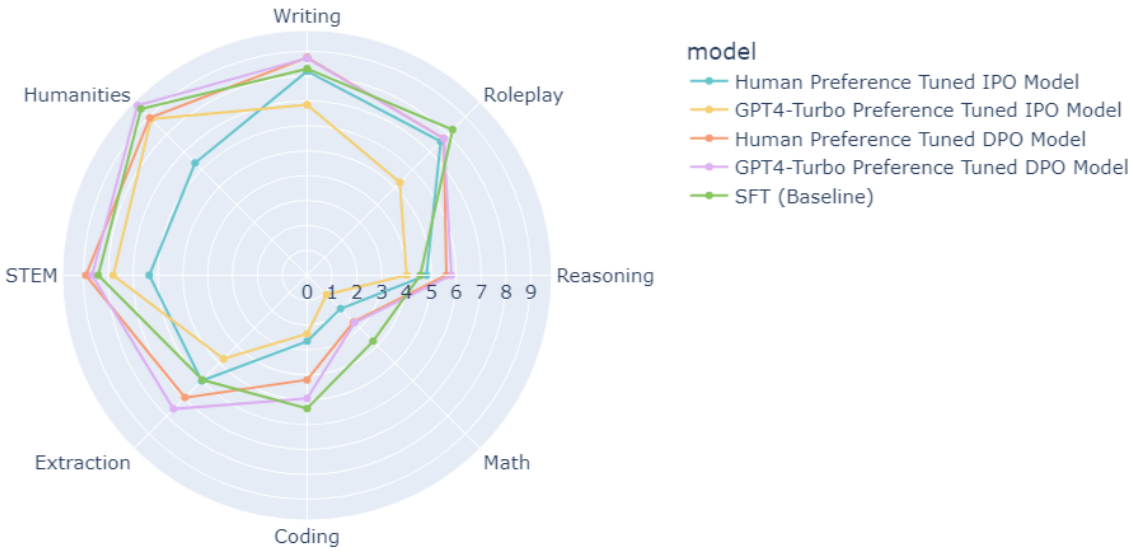

Figure 1: Analysis of GPT4 Ratings by Domains in MT Bench: DPO and IPO Models trained on GPT4-Turbo vs Human Preferences. Notably, the DPO model trained on GPT4-Turbo preferences excels over its counterpart trained on Human preferences in domains such as Coding, Extraction, Reasoning, and Humanities, while demonstrating competitive performance in other areas.

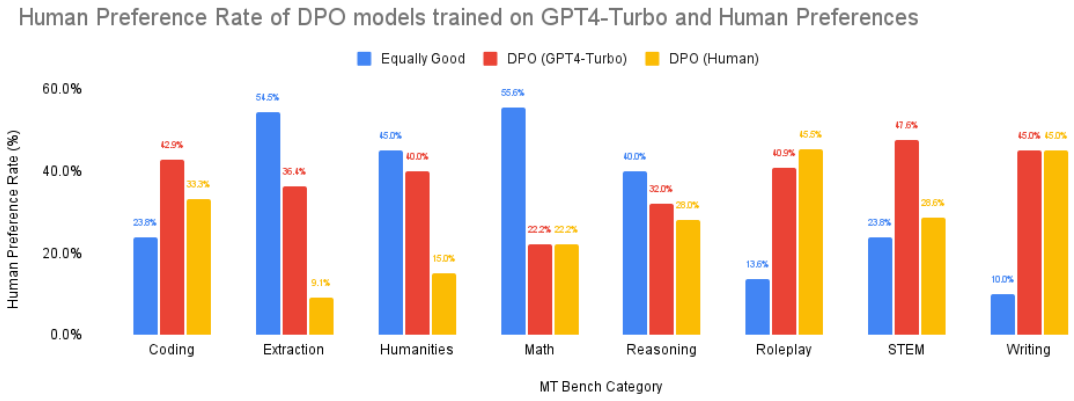

Figure 2: Analysis of Human Preference Rate by Domains in MT Bench: DPO Models trained on GPT4-Turbo vs. Human Preferences. Remarkably, the DPO model trained on GPT4-Turbo preferences demonstrates superior or comparable performance across all domains, with the exception of Roleplay.

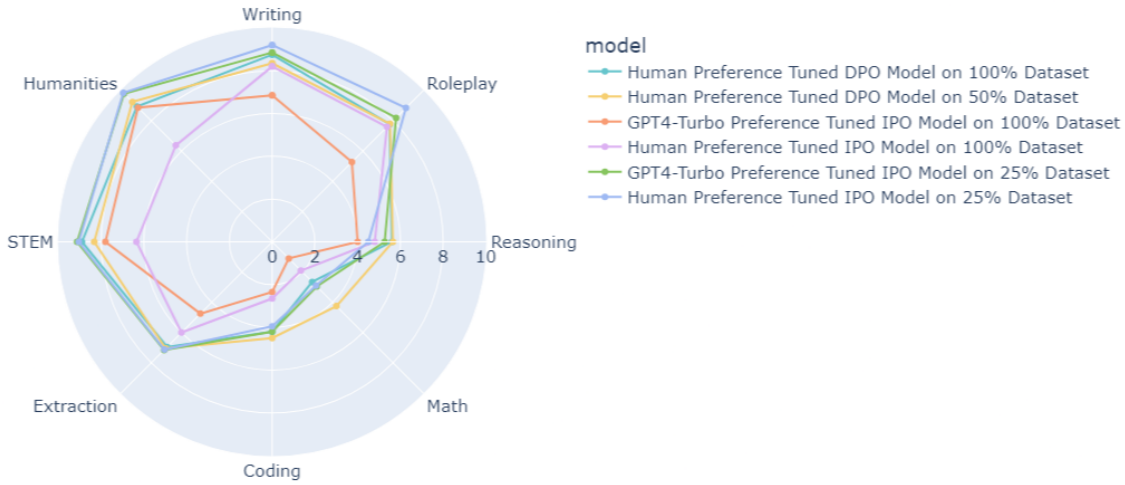

Figure 3: Analysis of GPT4 Ratings by Domains in MT Bench: DPO and IPO Models trained on different volumes of GPT4-Turbo and Human Preferences within the GAIR dataset. We present three comparisons that challenge the expected trends: IPO model trained on 25% versus 100% Human Preferences, IPO model trained on 25% versus 100% GPT4-Turbo Preferences, and DPO model trained on 50% versus 100% Human Preferences.

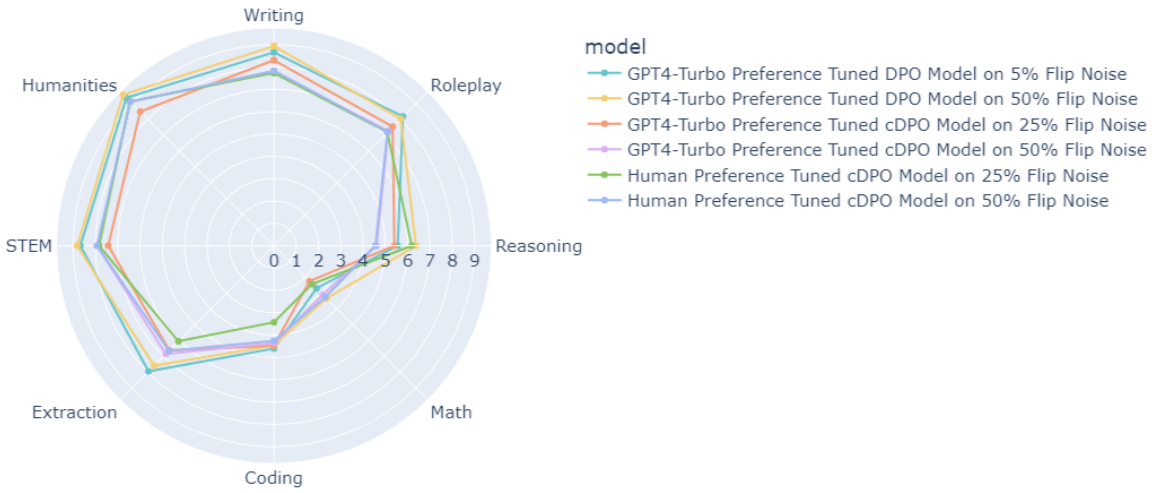

Figure 4: Analysis of GPT4 Ratings by Domains in MT Bench: DPO and cDPO Models trained on different proportions of flip noise induced in GPT4-Turbo and Human Preferences within the GAIR dataset. We present three noteworthy comparisons that challenge the expected trends: cDPO model trained on 25% versus 50% flip noise induced in Human Preferences, cDPO model trained on 25% versus 50% flip noise induced in GPT4-Turbo Preferences, and DPO model trained on 5% versus 50% flip noise induced in GPT4-Turbo Preferences.

Domain - GAIR (Train)	Domain - MT Bench (Test)	# Samples - GAIR (Train)
analyzing_general	Reasoning, Extraction, Writing, Roleplay	16
chitchat	Roleplay	239
code_correction_rewriting	Code	24
code_simplification	Code	1
counterfactual	Reasoning	52
explaining_code	Code	29
information_extraction	Extraction	30
keywords_extraction	Extraction	3
note_summarization	Extraction	1
question_generation	Reasoning	53
recommendation	Reasoning	45
solving_exam_question_with_math	Math	27
solving_exam_question_without_math	STEM, Humanities	39
text_simplification	Writing	7
text_to_text_translation	Writing	43
verifying_fact	Extraction	57
writing_cooking_recipe	Writing	47
writing_job_application	Writing	23
writing_marketing_materials	Writing	2
writing_personal_essay	Writing	29
writing_product_description	Writing	21
writing_social_media_post	Writing	10
writing_technical_document	Writing	13
creative_writing	Writing	275
instructional_rewriting	Writing	25
language_polishing	Writing	12
open_question	Writing	395
text_correction	Writing	14
title_generation	Writing	10
writing_advertisement	Writing	5
writing_email	Writing	79
writing_legal_document	Writing	17
writing_news_article	Writing	5
writing_presentation_script	Writing	12
writing_scientific_paper	Writing	6
writing_song_lyrics	Writing	41
functional_writing	Writing	195
paraphrasing	Writing	22
writing_blog_post	Writing	12
asking_how_to_question	Reasoning	100
classification_identification	Extraction	28
code_generation	Code	341
code_to_code_translation	Code	6
explaining_general	Reasoning	385
ranking	Reasoning	39
text_summarization	Extraction	93
brainstorming	Reasoning	165
data_analysis	Math	19
math_reasoning	Math, Reasoning	334
reading_comprehension	Reasoning	13
roleplay	Roleplay	131
value_judgement	Humanities	172
default	-	865
planning	Reasoning	75
seeking_advice	Roleplay	323

Table 6: Mapping between the domains represented in GAIR and MT Bench

Domain - MT Bench (Test)	# Samples - GAIR (Train)
Writing	1336
Reasoning	1277
Roleplay	924
Code	401
Math	380
Extraction	228
Humanities	211
STEM	39

Table 7: Data Volume in GAIR Corresponding to Domains in MT Bench

Data Volume (% Train Data)	Human	
	Loss = DPO (Tuned)	Loss = IPO (Tuned)
100%	7.184	7.113
75%	7.038	6.981
50%	7.181	6.722
25%	6.959	6.878

Table 8: Benchmarking performance of DPO and IPO models when trained with tuned β on different volumes of human preference data using MT Bench scores

Data Quality (% Flip Noise)	Human	
	Loss = DPO (Tuned)	Loss = cDPO (Tuned)
0%	7.184	7.184
5%	7.078	7.063

Table 9: Benchmarking performance of DPO and cDPO models when trained with tuned β on human preference datasets with different flip noise ratios using MT Bench scores

Algorithm	UltraFeedback	SHP
Baseline (SFT)	6.753	6.753
DPO	7.225	6.441

Table 10: Benchmarking DPO model performance on Ultrafeedback and SHP datasets using MT Bench scores

Data Volume (% Train Data)	UltraFeedback	SHP
100%	7.225	6.441
75%	7.419	6.438
50%	7.306	6.244
25%	7.384	6.122

Table 11: Benchmarking DPO model performance with varying sample proportions in Ultrafeedback and SHP datasets using MT Bench scores

Data Quality (% Flip Noise)	UltraFeedback	SHP
0%	7.225	6.441
5%	6.872	6.453
25%	6.691	6.272
50%	6.403	6.244
75%	5.928	5.664

Table 12: Benchmarking DPO model performance with varying flip noise in Ultrafeedback and SHP datasets using MT Bench scores

	UltraFeedback		SHP	
Data Volume (% Train Data)	Loss = DPO	Loss = IPO	Loss = DPO	Loss = IPO
100%	7.225	6.813	6.441	6.200
75%	7.419	5.853	6.438	6.469
50%	7.306	6.756	6.244	6.466
25%	7.384	6.344	6.122	6.419

Table 13: Benchmarking models with DPO and IPO loss functions across different Ultrafeedback and SHP dataset ratios using MT Bench scores

	UltraFeedback		SHP	
Data Quality (% Flip Noise)	**Loss = DPO**	**Loss = cDPO**	**Loss = DPO**	**Loss = cDPO**
0%	7.225	7.225	6.441	6.441
5%	6.872	6.484	6.453	6.425
20%	6.691	6.050	6.272	5.963
50%	6.403	6.325	6.244	6.403
75%	5.928	5.797	5.664	5.616

Table 14: Benchmarking models with DPO and cDPO loss functions across different Ultrafeedback and SHP flip noise ratios using MT Bench Scores.

Dataset	**# Train**	**# Val**	**Domains / Sub-Datasets**
Ultrafeedback	61135	2000	TruthfulQA, FalseQA, Evol-Instruct, UltraChat, ShareGPT
SHP	100000	1000	academia, anthropology, baking, carguys, culinary, docs, engineers, historians, hr, philosophy, physics, science, science fiction, social science, vet, change my view, explain like i'm five, legal advice
GAIR	5240	0	Chatbot Arena conversations
MT Bench	-	160	Writing, Roleplay, Extraction, Reasoning, Math, Coding, Knowledge I (STEM), and Knowledge II (humanities/social science)

Table 15: Overview of Datasets

RQ	**Description**	**Text**	**Output Score**
1	Prompt	You can see a beautiful red house to your left and a hypnotic greenhouse to your right, an attractive heated pink place in the front. So, where is the White House?	-
1	DPO Model trained on 100% of GPT4-Turbo preferences in GAIR dataset	The White House is not visible from this description as it is not mentioned. The description only describes other buildings and their colors.	3
1	DPO Model trained on 100% of Human preferences in GAIR dataset	The White House is likely located behind the red house and greenhouse, as it is the official residence and workplace of the President of the United States. The heated pink place in the front may be a separate building or structure on the same property.	2
1	Prompt	Benjamin went to a bookstore and purchased a variety of books. He bought 5 copies of a sci-fi novel, each priced at $20, 3 copies of a history book priced at $30 each, and 2 copies of a philosophy book for $45 each. What was the total cost of his purchases?	-
1	DPO Model trained on 100% of GPT4-Turbo preferences in GAIR dataset	The total cost of Benjamin's purchases is: 5 copies of sci-fi novel at $20 each = $100 3 copies of history book at $30 each = $90 2 copies of philosophy book at $45 each = $90 Total cost = $100 + $90 + $90 = $270	6
1	DPO Model trained on 100% of Human preferences in GAIR dataset	5 x $20 = $100 3 x $30 = $90 2 x $45 = $90 Total cost = $100 + $90 + $90 = $270 Therefore, the total cost of his purchases was $270.	4

Table 16: Examples of generated model outputs on MT Bench