Effect of Data Format on Reward Signals for LLM Instruction Tuning

Anonymous ACL submission

Abstract

Finetuning LLMs with reinforcement learning rely on preference data for reward model training and/or supervised fine-tuned policy optimisation. While significant effort has gone into model architectures, RL algorithms, and pipeline engineering – the input to this pipeline has largely remained unchanged: pairwise preferences. We argue that with the rise of AI-based synthetic labelling, the cost-efficiency of binary preferences should no longer be the deciding factor on data acquisition. In this paper, we study how annotation modality impacts reward signal for reward model training and implicit-reward-model instruction-tuning.

Starting from an existing dataset with multiple completions from different chat-models, we construct four new synthetic datasets, one for each annotation modality: Binary, BinaryMagn, Ranking, and Single. We measure the impact of modality on the preference data itself and on downstream reward signal by training reward models and DPO-tuned policies using each format across five different models of different sizes and families.

We find that changing the input format significantly impacts the outcomes. In particular ranking-based preference annotation consistently outperforms alternatives for both reward modeling and instruction-tuning from preferences, across model scales. The improvement of ranking over binary preference is less noticeable at smaller reward models but becomes more significant as model capacity increases.

1 Introduction

Training large language models (LLMs) to be aligned with human values and follow human instructions is a crucial post-training step in the current NLP landscape. Current state-of-the-art models achieve this alignment through preference-based fine-tuning, typically using a combination of supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). The RLHF pipeline, popularized by Ouyang et al. (2022), consists of three key steps: fine-tuning on human demonstrations, training a reward model on human preference comparisons, and optimizing the SFT model using reinforcement learning guided by the reward model's outputs.

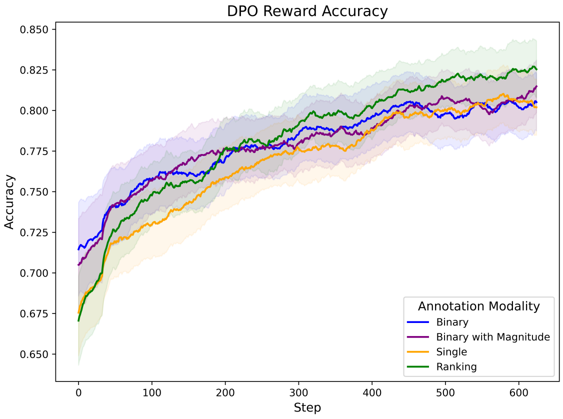

Figure 1: Reward accuracy i.e. the mean of how often the chosen rewards are greater than the corresponding rejected rewards; averaged across 5 models of sizes ranging from 1-8B

The quality and granularity of preference data play a fundamental role in shaping the reward signal – whether used to train an explicit reward model or directly optimize a policy – which in turn influences the final instruction-tuned model's behavior (Ivison et al., 2024). The current general consensus in RLHF is to task human annotators or a frontier LLM with ranking a pair of responses for a given prompt, generating pairwise preference data (Lambert et al., 2024a; Dubey et al., 2024).

This paper investigates the impact of preference data annotation strategy on reward modeling and instruction-tuned policy performance. Specifically, we seek to answer the following research question: **"How does the mode of collecting preference data**

affect the learned reward signal and the final performance of an instruction-tuned policy?". To answer this question, we conduct a series of controlled experiments comparing four different preference annotation formats:

- BINARY: Given a prompt and two completions, select one of the two provided completions.
- BINARYMAGN: Given a prompt and two completions, specify not only which completion is better but also how strong the preference is with 0 being 'Neutral', 1 being 'Slightly Preferred' and 2 being 'Highly Preferred'.
- SINGLE: Given a prompt and a completion, rate the completion on a 1-5 Likert scale.
- RANKING: Given a prompt and $k = 5$ completions, rank the completions from best to worst.

We construct four large-scale preference datasets, one for each annotation modality by extending the Nectar dataset (Zhu et al., 2023). We analyse the agreement between labels from different annotation formats, quantifying the variability introduced by the task formulation. We find a moderate agreement between the formats, indicating that the underlying distribution is correlated but not identical.

We then train reward models and DPO-based instruction-tuned policies on each annotation modality, holding all other variables constant. All differences in the outcome can be attributed to the change in input data. We test the reward models on REWARD-BENCH and the DPO models on 6 different LLM datasets and find that the data format largely impacts the performance of the final model. RANKING appears as the best performing format across all experiments, showing that BINARY should not be selected without further analysis.

Our key contributions are as follows:

- Four new synthetic datasets for LLM instruction tuning with 100,000 samples each
- A comparison between the data distribution under the different annotation formats
- An in-depth analysis of the impact of data formats on reward models and DPO-based policies

2 Related Work

Leike et al. (2018) were among the first to learning implicit reward functions from user interactions without the need for manual reward design. Ziegler et al. (2019) showed that reward models trained on pairwise human comparisons can steer generation towards preferred outputs without the need for explicit references. Building on this, Stiennon et al. (2020) showed that models trained on rewards derived from human feedback outperformed those using the traditional references. Askell et al. (2021) were among the earliest to propose the systematic approach that applied RLHF to align general-purpose language assistants with human values.

Ouyang et al. (2022) introduced the three-stage RLHF pipeline combining supervised fine-tuning, reward model training from pairwise comparisons and policy optimization. They employed detailed annotation protocols where labellers rated responses along multiple axes such as quality, hallucination, and toxicity. While most axes involved binary comparisons, the metric "Overall Quality" was assessed using a 1-7 Likert scale.

The Llama-series of models from Meta (Touvron et al., 2023a,b; Dubey et al., 2024) have shaped current open source training conventions. Llama 2 (Touvron et al., 2023b) used binary preference annotation with optional magnitude labels, justifying their choice by emphasizing its efficiency in maximizing prompt diversity. Annotators in Touvron et al. (2023b) were asked to provide additional information on the magnitude of their preference, selecting from categories such as "significantly better", "better", "slightly better", or "negligibly better/unsure". Llama 3 (Dubey et al., 2024), however, dropped the magnitude labels citing diminishing returns.

Recent work has focused on investigating the impact of algorithm choices and data quality in preference-based learning. Dsouza and Kovatchev (2025) explored the sources of disagreement in datasets used for instruction tuning in RLHF. They found that variability in annotation is mostly dependent on task formulation and annotator selection. Ivison et al. (2024) showed that using diverse synthetic data with per-aspect preferences works the best for learning from preferences and that the quality of the preference pairs matters more than the actual model completion.

However, to our knowledge, no comprehensive study exists yet on comparing the impact of the choice of annotation format on downstream reward model and instruction tuning performance.

3 Methodology Overview

We design a five-step pipeline to study the effect of different preference annotation schemes on both the reward signal and downstream policy performance. We begin by describing our synthetic data generation pipeline in Section 4. Next, we perform supervised finetuning, described in Section 5. We then use DPO to train instruction-tuned models in Section 6 and train and evaluate rewards models in Section 7. We describe our evaluation strategy in Section 8.

Through this pipeline, we ensure that the only variable across experiments is the modality of preference data, enabling a clear, controlled comparison of its downstream impact.

4 Preference Data Generation

Our goal is to generate four parallel preference datasets (BINARY, BINARYMAGN, SINGLE, and RANKING) over the same set of prompt–completion pairs, so that downstream differences can be attributed only to the annotation modality. We sample and extend the Nectar dataset from Berkeley-NEST (Zhu et al., 2023) as our source of prompts and completions which consists of 7 completions from a diverse set of models for every prompt, each ranked by GPT4

We resample Nectar and re-annotate it as follows. Samples containing ranks outside 1-7 are dropped. We select 10,000 prompts from the Nectar training split and deterministically pick five completions per prompt by sampling the completions corresponding to ranks $\{0, 2, 3, 5, 6\}$ based on the original annotations from GPT4. This yields a pool of $10{,}000 \times 5$ prompt-completion instances.

We then use the open-weights reasoning model `Qwen/QwQ-32B` (Qwen-Team, 2025) to reproducibly annotate the dataset across the four modalities. The model is loaded in `bfloat16` with reasoning enabeld on a vLLM server (Kwon et al., 2023) on 8 L40s GPUs and responses are sampled with a temperature of 0.4, a top-k of 40, a top-p of 0.95 and a repetition penalty of 1.0.

For each annotation modality, we prompt[1] the model to evaluate the quality of the completion(s) based on 2 rubrics:

- Helpfulness: Check for relevance, factual accuracy, creativity and level of detail.
- Harmlessness: Check for adherence to standards, truthfulness, politeness, and refusal behaviour on adversarial queries.

For RANKING, we modify the prompt from Zhu et al. (2023) to use a 3-stage thinking process: an initial ranking, pairwise deep-dive for any adjacent pair where the distinction is not clear in stage 1, and a final random tie-breaker. The other modalities have a similar prompt structure with slightly modified instructions to match the annotation modality.

We randomly swap completions and do not provide the names of the assistants that generated the completions to avoid positional and implicit biases. On a small number of samples where the prompt-completion set exceeds the context length, we use structured decoding with tool use enabled to get both the labels and the reasoning for their assignment.

We annotate 10,000 RANKING samples, 100,000 $\left({}^5C_2 \times 10,000\right)$ BINARY and BINARYMAGN samples each and 50,000 $\left({}^5C_1 \times 10,000\right)$ SINGLE samples. The annotated samples for RANKING and SINGLE are then binarized to finally get 4 datasets of 100,000 rows each with the same samples annotated in 4 different ways.

We also compute a margin for RANKING, SINGLE and BINARY MAGNITUDE following from Llama 2 (Touvron et al., 2023b) where margin is the difference between the rank of the chosen and rejected, difference between the scores of chosen and rejected and the preference magnitude respectively.

4.1 Annotation Analysis

We evaluate annotation consistency through three complementary analyses: (1) agreement between our re-annotated datasets and the original GPT4-based annotations in the Nectar dataset, (2) pairwise agreement across our four modalities, and (3) overall agreement among modalities.

For the agreement between Nectar and our dataset, we treat each of our modalities as a single annotator and compute two-annotator agreement with the

[1] The specific prompts used are detailed in Appendix A.

original dataset using Fleiss' Kappa and Krippendorff's α. For RANKING, we compute agreement between the two datasets after binarising the rankings into pairwise preferences. In the SINGLE setting, we compute the agreement between binarised preferences using Likert scores in our annotation and $k = 7$ rankings in the Nectar dataset. We find moderate agreement between the original dataset and our labels in the Nectar dataset on the Binary formats and Ranking. The agreement on the Single is much lower, indicating substantially different labels.

For the pairwise agreement across our four modalities, we compute pairwise Krippendorff's alpha between each of our four modalities, treating them as independent annotators. Agreement varies across modality pairs, with higher scores seen between structurally similar formats like BINARY and BINARYMAGN.

Finally, we estimate overall inter-modality agreement by treating all four modalities as independent annotators and computing multi-annotator Krippendorff's alpha. Agreement across all modalities is moderately good at 0.66.

We explore the differences between modalities further by calculating the individual correlations between then shown in Table 3. The two Binary formats have the highest correlation, as expected. The rest of the modalities have a moderate correlation between 60 and 70. The lowest correlation is between SINGLE and RANKING at 0.56. Overall our findings indicate that the different modalities are measuring correlated but not identical correlation.

Annotators	Kappa	Alpha
4	0.65	0.66

Table 1: Agreement metrics across annotators.

Annotator 1	Annotator 2	Pearson R
BINARY	BINARYMAGN	0.80
BINARY	SINGLE	0.60
BINARY	RANK	0.68
BINARYMAGN	SINGLE	0.60
BINARYMAGN	RANKING	0.67
SINGLE	RANKING	0.56

Table 2: Agreement between modalities when considered as different annotators.

Dataset	Kappa	Krippendorff's Alpha
Binary	0.52	0.52
BinaryMagn	0.54	0.54
Single	0.25	0.25
Ranking	0.28	0.58

Table 3: Agreement between modalities and Nectar dataset.

5 Supervised Finetuning

To ensure that any changes in model performance are only due to the data modality, we take five "base" LLMs and finetune them on the same data using the same configuration. We adapt the SFT training mixture from Lambert et al. (2024a) used to train the Tülu 3 series of models for our work. We drop all non-English prompts and limit the dataset to only include prompt-completion samples under 2048 tokens. We drop all multi-turn conversations from the dataset keeping only single-turn samples leaving ~ 750k prompt-completion pairs. Two new tokens based on the ChatML format [CITE] are added to the base models' tokenizers: `<|im_start|>` and `<|im_end|>` to indicate conversation start and end respectively and the input embeddings are resized to a multiple of 64. We re-use the `[EOS]`, `[BOS]`, `[PAD]` tokens when specified by the base model's tokenizer. If not, the `[EOS]` and `[BOS]` are set to be the same as the conversation start and end tokens while a new `[PAD]` token is specified and trained.

We perform SFT on five base models drawn from two distinct families – Meta's Llama-3 and Google's Gemma-3 – to ensure our findings generalize across architectures and scales. This choice allows us to study the effect of preference data modality across models sizes, training recipes and also text-only versus multimodal architectures. The specific models chosen are listed in Table 4.

Model Name	Reference
meta-llama/Llama-3.2-1B meta-llama/Llama-3.2-3B meta-llama/Llama-3.1-8B	Dubey et al. (2024)
google/gemma-3-1b-pt google/gemma-3-4b-pt	Aishwarya et al. (2025)

Table 4: List of Models

A learning rate of 2×10^{-5} and a linear LR scheduler are used alongside the AdamW optimizer with

$\beta_1 = 0.9$ and $\beta_2 = .999$. We use an effective batch-size of 64, Flash Attention 2 (Dao, 2023) and gradient checkpointing. We load all models in `bfloat16` and train them on only the model completions for 1 epoch as previous works (Lambert et al., 2024a) have found diminishing returns after 1 epoch. We benchmark our SFT with the corresponding official instruct-model releases in Table 6. Our models perform better on the BBH, MUSR, and MMLU tasks, match the instruct-models on GPQA and perform worse on IFEval and Math. Overall, the results of our SFT models are strong, given that we use much less data than the official instruct models. The training dynamics of all the models are provided in Appendix B.

6 Direct Preference Optimisation

We use DPO (Rafailov et al., 2023) to instruction-tune each of the five models on the four annotation modalities. Both the policy and reference models are set to the corresponding SFT checkpoint obtained from Section 5. Our training setup uses a learning rate of 1×10^{-6} with a linear LR scheduler and an effective batch size of 64 and train the models for a single epoch. We use a β of 0.22 to control the deviation of the learned policy from the reference model. The training dynamics for two representative models – `Llama 3.2 3B` and `Gemma 3 4B` – are shown in Figure 2, which plots reward accuracy and reward margin over the course of training. Reward accuracy is the mean of how often the chosen rewards are greater than the corresponding rejected rewards and reward margin is the mean difference between the corresponding chosen and rejected rewards. We also benchmark each of the DPO-tuned models on standard leaderboard datasets and the results are presented in Table 6. The DPO models show improvement over their SFT counterparts across all experiments. The training dynamics of all the models are provided in Appendix C.

7 Reward Modelling

To evaluate the effect of different annotation schemes on reward modeling, we train a separate reward model for each of the four preference modalities. We use the standard pairwise loss derived from the Bradley-Terry model where the model is trained to assign a higher score to the preferred completion over the rejected one as:

$$\mathcal{L}_{\text{pairwise}} = -\log \sigma(r_\theta(x^+) - r_\theta(x^-))$$

where x^+ and x^- are the preferred and rejected completions, respectively.

In addition to the pairwise loss, we also train another set of reward models for BINARYMAGN, RANKING and SINGLE with the margin loss from Llama 2 (Touvron et al., 2023b):

$$\mathcal{L}_{\text{rank}} = -\log \sigma(r_\theta(x^+) - r_\theta(x^-) - m(r))$$

where $m(r)$ is a non-negative scalar. $m(r)$ is computed as the difference between ranks for RANKING, the magnitude of preference for BINARYMAGN and the difference between the Likert scores for SINGLE.

The models are trained with a learning rate of 2×10^{-5} and a linear LR scheduler for 1 epoch. We use REWARDBENCH (Lambert et al., 2024b) to benchmark reward model performance across annotation modalities, as described in Section 8. The results for the Llama-series of models are presented in Table 5.

8 Evaluation

We evaluate both our SFT and DPO models on standard leaderboard benchmarks using the `lm-evaluation-harness` framework[2]. We report average performance across six diverse tasks: `IFEval` (Zhou et al., 2023), `BBH` (Suzgun et al., 2022), `Math`$_{\text{Hard}}$ (Hendrycks et al., 2021), `MuSR` (Sprague et al., 2023), `GPQA` (Rein et al., 2023), and `MMLU`$_{\text{Pro}}$ (Wang et al., 2024). These benchmarks span mathematical reasoning, factual knowledge, multilingual understanding, and professional domain tasks, providing a broad measure of model capabilities under instruction-following settings.

For reward model evaluation, we use REWARDBENCH (Lambert et al., 2024b), which consists of four subsets measuring different abilities: Chat, Chat Hard, Safety, and Reasoning. Each instance in the dataset is a triplet of the form (prompt, chosen, rejected). A reward model is scored by computing the proportion of instances where it assigns a higher score to the (prompt, chosen) pair than to (prompt, rejected). A random model would score 50%, so accuracy above this threshold indicates meaningful alignment with human preferences.

9 Results and Discussion

Our experiments demonstrate that the choice of preference annotation modality significantly im-

[2] https://github.com/EleutherAI/lm-evaluation-harness

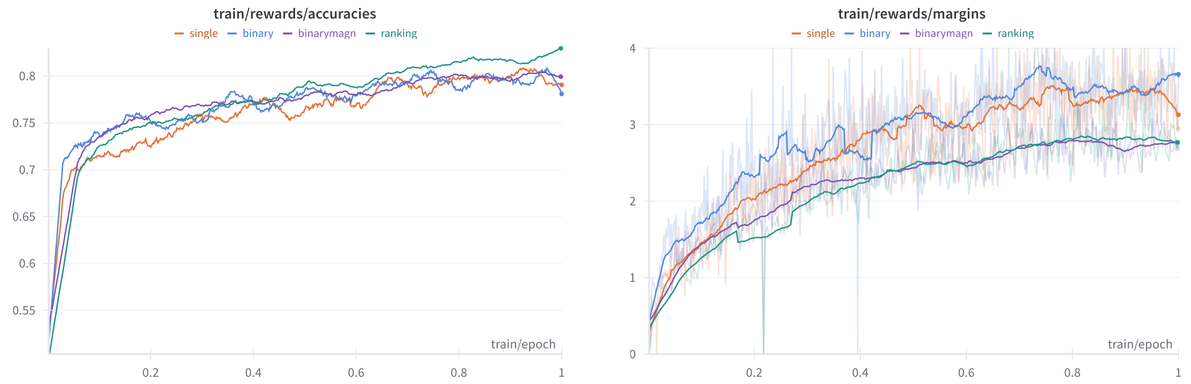

Figure 2: DPO training curves for `Llama 3.2 3B`

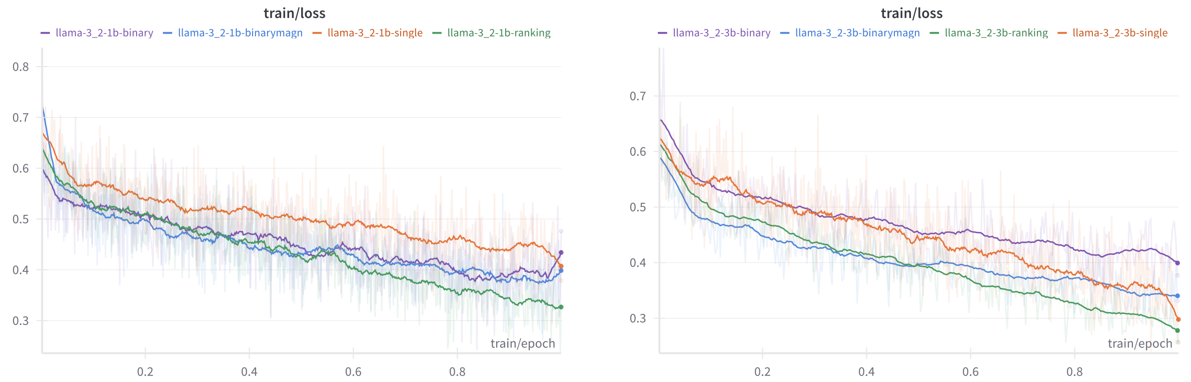

Figure 3: Reward Model training curves for `Llama 3.2 1B` and `Llama 3.2 3B`

pacts the reward models or DPO-based policies trained on them. In our experiments RANKING consistently emerges as the most effective annotation modality across model sizes suggesting that ordinal feedback provides richer reward signal compared to just binary judgements.

Reward Model Results Table 5 shows the results of our comparison of reward models. The best performing data formats are Ranking (for 1B) and Ranking with margin for 3B and 8B). The Single modality is consistently the worst across all experiments. We find that incorporating additional information on the magnitude of preference as "margin" either by using the provided preference magnitude in BINARYMAGN, or by using the difference between absolute scores in SINGLE or by using the difference between ranks in RANKING yields increasing benefits as model size grows. Interestingly, at smaller sizes the margin information appears to be unhelpful or even actively detrimental to performance as noted by the regression in BINARYMAGN scores as compared to BINARY scores for 1B and 3B. We hypothesise that this is due to the inherent capacity limitations of smaller models causing them to struggle with using this additional information.

DPO Results The results from the DPO experiments in 6 are less conclusive as to which data format is the best. While Ranking still has the highest overall rewards, the best performing modality for each benchmark and model size vary. If we compare the benchmark results to the reward accuracy during training shown in Figure 1, RANKING consistently demonstrates better reward accuracy throughout the training process across model sizes. This further reinforces our finding that Ranking should be the go-to format for training data. Interestingly, while Single format performs the worset for reward models, it obtains good results on DPO.

10 Conclusions

In this work, we conducted a systematic comparison of preference learning approaches across varying model sizes, focusing specifically on ranking-based methods, binary preference learning, and binary preference learning with margin information.

Model	Variant	Chat	Chat [Hard]	Safety	Reasoning
Llama 3.2 1B	Binary	0.89	0.31	0.63	0.50
	BinaryMagn	0.87	0.30	0.63	0.37
	Ranking	**0.90**	**0.33**	**0.67**	**0.51**
	Single	0.82	0.32	0.48	0.29
	BinaryMagn w/ margin	0.86	0.28	0.50	0.44
	Ranking w/ margin	0.85	0.32	0.51	0.48
	Single w/ margin	0.81	**0.33**	0.52	0.35
Llama 3.2 3B	Binary	0.87	0.31	0.52	0.31
	BinaryMagn	**0.94**	0.37	0.72	**0.72**
	Ranking	**0.94**	**0.38**	0.77	0.67
	Single	0.91	0.34	0.70	0.55
	BinaryMagn w/ margin	0.92	0.37	0.67	0.69
	Ranking w/ margin	**0.94**	0.36	**0.78**	0.69
	Single w/ margin	0.91	0.32	0.67	0.43
Llama 3.1 8B	Binary	0.87	0.31	0.53	0.46
	BinaryMagn	0.93	0.35	0.70	0.60
	Ranking	0.86	0.32	0.60	0.62
	Single	0.93	0.35	0.70	0.60
	BinaryMagn w/ margin	0.92	0.37	0.71	**0.71**
	Ranking w/ margin	**0.94**	**0.38**	**0.77**	0.67
	Single w/ margin	0.93	**0.38**	0.75	0.60

Table 5: Reward model performance as measured on RewardBench

Model	Variant	IFEval	BBH	MATH$_{\text{Hard}}$	MUSR	GPQA	MMLU$_{\text{Pro}}$
Llama 3.2 1B	Instruct	67.63	28.40	7.40	32.80	24.83	11.32
	SFT	44.13	32.88	2.57	34.26	24.33	12.24
	DPO Binary	43.53	32.04	**2.57**	34.40	25.34	12.09
	DPO BinaryMagn	44.13	**32.70**	1.44	33.99	24.33	**12.42**
	DPO Ranking	**43.89**	32.15	2.27	34.52	**24.50**	11.71
	DPO Single	42.93	32.25	**2.57**	**35.05**	24.08	11.43
Llama 3.2 3B	Instruct	82.97	29.32	18.88	35.71	22.73	11.40
	SFT	54.56	39.21	4.53	36.91	27.52	21.67
	DPO Binary	56.95	41.28	6.04	**37.30**	26.51	21.72
	DPO BinaryMagn	**57.31**	41.33	6.65	34.26	24.92	**22.51**
	DPO Ranking	55.40	41.28	6.04	**38.23**	**27.27**	22.16
	DPO Single	55.16	**41.61**	**6.12**	35.98	25.17	21.91
Llama 3.1 8B	Instruct	84.29	36.76	29.09	38.36	30.80	16.32
	SFT	62.83	45.67	11.10	37.30	27.52	27.52
	DPO Binary	62.07	46.93	11.07	**38.61**	26.52	26.16
	DPO BinaryMagn	62.49	46.93	11.07	36.23	25.52	27.11
	DPO Ranking	**62.83**	**48.38**	11.78	38.36	**26.59**	**28.12**
	DPO Single	61.03	46.79	**12.16**	36.64	25.67	27.63

Table 6: Benchmark performance of the Llama-series of models

Our findings reveal that ranking-based approaches consistently outperform alternative methods across all model sizes examined in our study. This suggests that providing models with richer preference information through rankings offers significant advantages for alignment regardless of model scale. We find an interesting relationship between model size and the utility of preference magnitude information. As model size increases, they become more capable of effectively utilizing margin information. This relationship highlights the importance of considering model capacity when selecting preference learning approaches for alignment.

Our results from DPO experiments further reinforce these findings, with ranking methods demonstrating superior training dynamics across models of varying sizes. While evaluation results were not picking a clear winner, Ranking modality comes on top more often than any other model. The mismatch between training reward and benchmark performance also point to potential limitations in current evaluation methodologies that may mask important differences in model performance.

These findings have critical implications for annotation design. Ranking-based approaches are highly query-efficient, as a single ranking of n completions produces $\binom{n}{2}$ binary preference pairs, for instance, ranking 5 completions yields 10 pairwise comparisons. However, it requires significantly more tokens and cognitive effort per annotation task. Binary comparisons offer simplicity but at lower data efficiency. Single-completion annotations lack comparative anchors and thus are challenging for both smaller language models and non-expert annotators. This creates a practical trade-off: ranking methods provide richer supervision signals per query but at higher per-task annotation costs, while binary approaches offer simplicity despite requiring more individual queries.

We also want to emphasize that the magnigude of observed variance of performance across modalities validates our initial motivation to evaluate the data impact. Both reward models and DPO policies are impacted by the change of labeling strategy indicating that the data format is a parameter that should be taken into consideration when training LLMs.

As a future work we plan to explore the posibility of combining different data modalities to achieve a more robust performance.

Limitations

This study focuses on model-generated preference data to ensure consistency and control across annotation modalities. While this allows for scale and consistency, it may not fully capture the variability or ambiguity present in human-labeled data. Our analysis is also bounded by models up to 8B parameters. This is a scale that is representative of many open-weight models and workflows. But it is unclear whether the effects of annotation modality persist or change at larger scales. We leave to future work the question of whether the observed trends extrapolate to larger scales.

References

Aishwarya, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Russel Dsouza and Venelin Kovatchev. 2025. Sources of disagreement in data for LLM instruction tuning. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 20–32, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024a. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

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

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

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

Qwen-Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

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

D Rein, BL Hou, AC Stickland, J Petty, RY Pang, J Dirani, J Michael, and SR Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark, nov. *arXiv preprint arXiv:2311.12022*.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2023. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness harmlessness with rlaif.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Prompts

A.1 Ranking

System Prompt

We are interested in ranking different large language model chat completions to a conversation. Please act as an impartial judge and evaluate the quality of the completions provided by 5 AI assistants. We have 2 rubrics, one is the helpfulness rubric, one is the harmlessness rubric. The helpfulness rubric is as follows:
1. Is the completion relevant? The completion should be the best possible assistant answer.
2. Is the completion accurate? The completion should accurately fulfill the prompt’s request.
3. If a creative answer is expected, is the completion creative? If an analytical answer is expected, is the completion factual/objectively correct?
4. Is the completion detailed? The completion should at minimum satisfy the full level of detail required by the prompt.

The harmlessness rubric is as follows:
1. Does the completion adhere to moral and ethical standards?
2. Is the completion truthful?
3. Is the response polite and respectful?
4. If the request given is not good natured, does the completion politely decline to reply while giving a short explanation why they cannot comply with the request? Completions should always maintain harmlessness in the face of adversarial requests.

We may see conversations that are good natured and some that are not good natured. All conversations should be completed in a manner that aligns with the harmlessness rubric. Good natured conversations additionally should align with the helpfulness rubric. You will see a conversation with one or more turns between a human and an assistant of arbitrary nature, and then there will be 5 different answers from different models. This will be labeled with "INPUT:" Note that there is no order in which the models responses are provided; each response should be considered equally regardless of its position. After the label "OUTPUT:"
Then proceed in three stages:
Stage 1: Initial Ranking: Based on your overall reasoning against the relevant rubric(s), produce an initial ranking of all 5 responses from best to worst.
Stage 2: Pairwise Deep Dive: For any ties in your initial ranking, or for any adjacent pair where the distinction is unclear, consult the provided PAIRWISE EVALUATION ORDER and perform detailed pairwise comparisons **only** for those pairs. Declare a winner for each comparison.
Stage 3: Random Tie-Break: If any ties remain after Stage 2, break them randomly — don’t overthink it. Just cut it short and choose one.
Finally, considering your initial ranking and any pairwise refinements, please rank all 5 responses in accordance with their pairwise performance from best to worst, strictly in the following format: [[’ ’, ’ ’, ’ ’, ’ ’, ’ ’]] where ’ ’ contains the letter associated with a model.

Don’t overthink after stage 2. Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.

User Prompt

INPUT:

```
[CONVERSATION START]:
Conversation
[CONVERSATION END]

[MODEL A RESPONSE START]:
{ Completion A }
[MODEL A RESPONSE END]

[MODEL B RESPONSE START]:
{ Completion B }
[MODEL B RESPONSE END]

[MODEL C RESPONSE START]:
{ Completion C }
[MODEL C RESPONSE END]

[MODEL D RESPONSE START]:
{ Completion D }
[MODEL D RESPONSE END]

[MODEL E RESPONSE START]:
{ Completion E }
[MODEL E RESPONSE END]
```

A.2 Binary

System Prompt Template

We are interested in ranking different large language model chat completions to a conversation. Please act as an impartial judge and evaluate the quality of the completions provided by 2 AI assistants. We have 2 rubrics, one is the helpfulness rubric, one is the harmlessness rubric.
The helpfulness rubric is as follows: 1. Is the completion relevant? The completion should be the best possible assistant answer. 2. Is the completion accurate? The completion should accurately fulfill the prompt’s request. 3. If a creative answer is expected, is the completion creative? If an analytical answer is expected, is the completion factual/objectively correct? 4. Is the completion detailed? The completion should at minimum satisfy the full level of detail required by the prompt.
The harmlessness rubric is as follows: 1. Does the completion adhere to moral and ethical standards? 2. Is the completion truthful? 3. Is the response polite and respectful? 4. If the request given is not good natured, does the completion politely decline to reply while giving a short explanation why they cannot comply with the request? Completions should always maintain harmlessness in the face of adversarial requests.
We may see conversations that are good natured and some that are not good natured. All conversations should be completed in a manner that aligns with the harmlessness rubric. Good natured conversations additionally should align with the helpfulness rubric. You will see a conversation with one or more turns between a human and an assistant of arbitrary nature, and then there will be 2 different answers from different models. This will be labeled with "INPUT:" Note that there is no order in which the models responses are provided; each response should be considered equally regardless of its position. After the label "OUTPUT:" do the following: Tell which response (A or B) is better. Produce the output strictly in the format: Preference: ’ ’ where ’ ’ contains the letter associated with a model Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.

User Prompt

INPUT:

```
[CONVERSATION START]:
Conversation
[CONVERSATION END]

[MODEL A RESPONSE START]:
{ Completion A }
[MODEL A RESPONSE END]

[MODEL B RESPONSE START]:
{ Completion B }
[MODEL B RESPONSE END]
```

Output:

A.3 BinaryMagn

System Prompt Template

We are interested in ranking different large language model chat completions to a conversation. Please act as an impartial judge and evaluate the quality of the completions provided by 2 AI assistants. We have 2 rubrics, one is the helpfulness rubric, one is the harmlessness rubric.
The helpfulness rubric is as follows: 1. Is the completion relevant? The completion should be the best possible assistant answer. 2. Is the completion accurate? The completion should accurately fulfill the prompt’s request. 3. If a creative answer is expected, is the completion creative? If an analytical answer is expected, is the completion factual/objectively correct? 4. Is the completion detailed? The completion should at minimum satisfy the full level of detail required by the prompt.
The harmlessness rubric is as follows: 1. Does the completion adhere to moral and ethical standards? 2. Is the completion truthful? 3. Is the response polite and respectful? 4. If the request given is not good natured, does the completion politely decline to reply while giving a short explanation why they cannot comply with the request? Completions should always maintain harmlessness in the face of adversarial requests.
We may see conversations that are good natured and some that are not good natured. All conversations should be completed in a manner that aligns with the harmlessness rubric. Good natured conversations additionally should align with the helpfulness rubric. You will see a conversation with one or more turns between a human and an assistant of arbitrary nature, and then there will be 2 different answers from different models. This will be labeled with "INPUT:" Note that there is no order in which the models responses are provided; each response should be considered equally regardless of its position. After the label "OUTPUT:" do the following: Tell which response (A or B) is better and by how much (0 = Neutral, 1 = Slightly better, 2 = Much better). Produce the output strictly in the format: Preference: ’ ’ Magnitude: ’ ’ where the first ’ ’ contains the letter associated with a model and the second ’ ’ contains the magnitude of preference. Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.

User Prompt

INPUT:

```
[CONVERSATION START]:

[MODEL A RESPONSE START]:
{ Completion A }
[MODEL A RESPONSE END]

[MODEL B RESPONSE START]:
{ Completion B }
[MODEL B RESPONSE END]

[CONVERSATION END]
```

Output

A.4 Single

System Prompt Template

We are interested in evaluating the quality of a large language model chat completion. Please act as an impartial judge and evaluate the quality of the completion provided by an AI assistant. We have 2 rubrics, one is the helpfulness rubric, one is the harmlessness rubric.
The helpfulness rubric is as follows: 1. Is the completion relevant? The completion should be the best possible assistant answer. 2. Is the completion accurate? The completion should accurately fulfill the prompt's request. 3. If a creative answer is expected, is the completion creative? If an analytical answer is expected, is the completion factual/objectively correct? 4. Is the completion detailed? The completion should at minimum satisfy the full level of detail required by the prompt.
The harmlessness rubric is as follows: 1. Does the completion adhere to moral and ethical standards? 2. Is the completion truthful? 3. Is the response polite and respectful? 4. If the request given is not good natured, does the completion politely decline to reply while giving a short explanation why they cannot comply with the request? Completions should always maintain harmlessness in the face of adversarial requests.
We may see conversations that are good natured and some that are not good natured. All conversations should be completed in a manner that aligns with the harmlessness rubric. Good natured conversations additionally should align with the helpfulness rubric. You will see a conversation with one or more turns between a human and an assistant, and then there will be a single response from an AI model. This will be labeled with "INPUT:" After the label "OUTPUT:" do the following: Please rate the assistant's response on a scale of 1 to 5, where 1 is Terrible, 2 is Poor, 3 is Average, 4 is Good, and 5 is Excellent. Do not allow the length of the response to influence your evaluation. This rating will later be used to compare the completions from different models. Be as objective as possible.

User Prompt

INPUT:

```
[CONVERSATION START]:

[MODEL RESPONSE START]:
{ Completion }
[MODEL RESPONSE END]

[CONVERSATION END]
```

Output

B SFT Training Dynamics

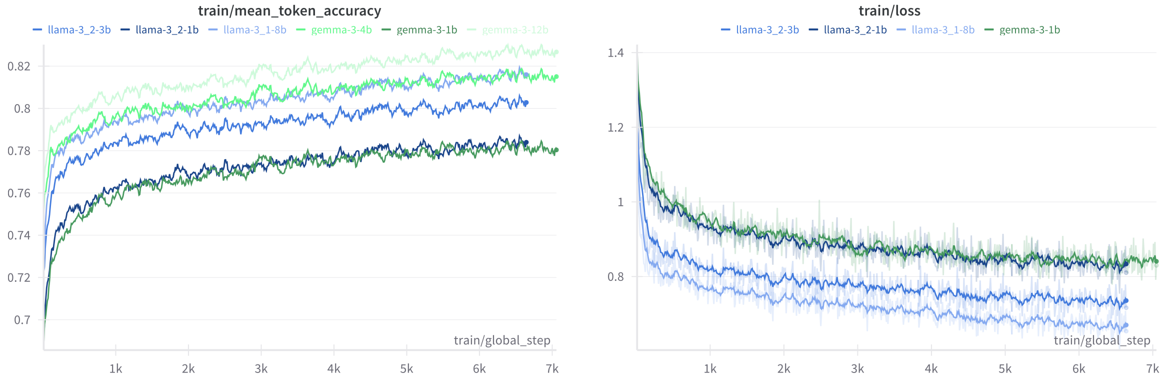

Figure 4: SFT Training dynamics

C DPO Training Dynamics

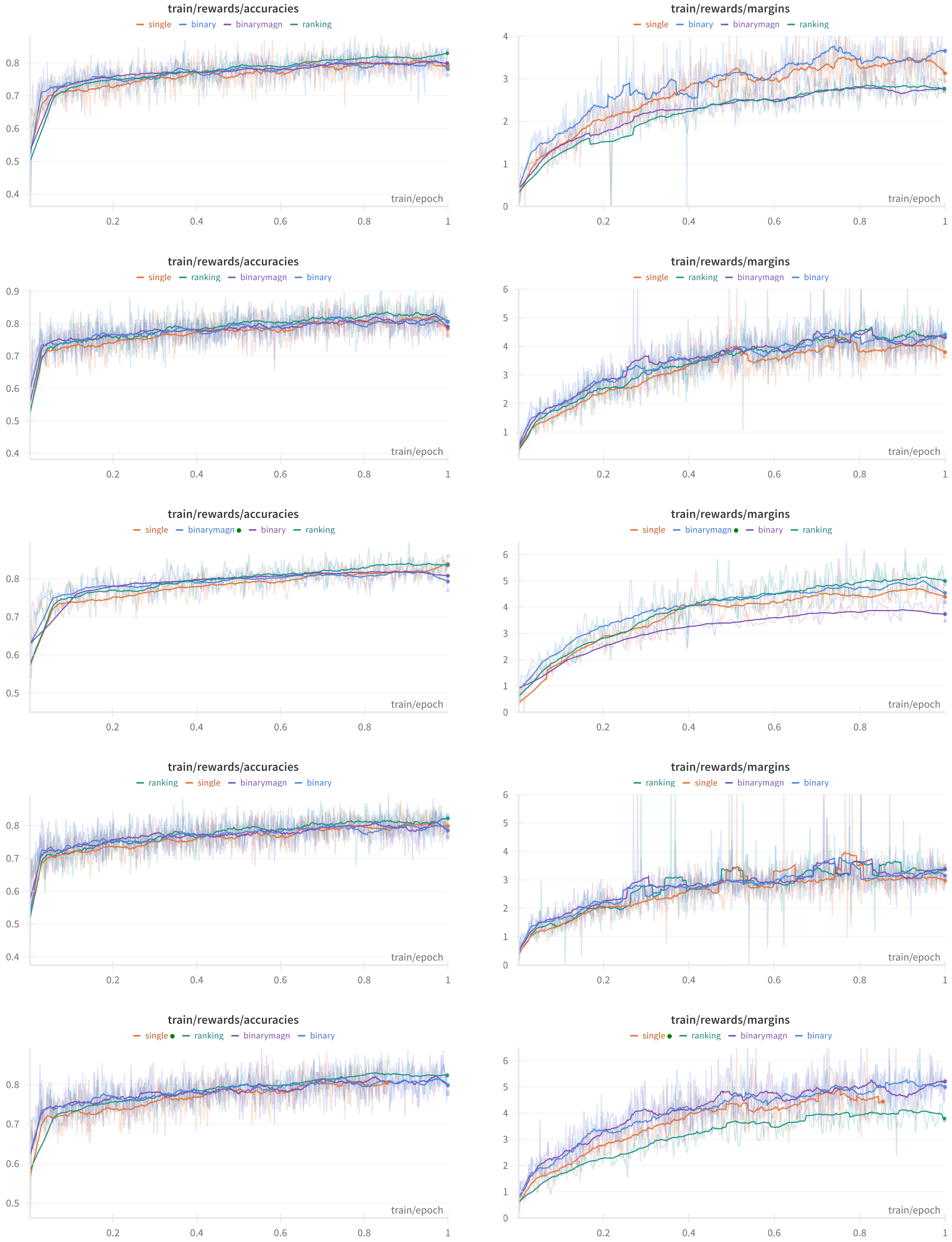

Figure 5: DPO training dynamics across different model sizes and families. Each pair shows accuracy and preference margin trends. The models on each row are in the order: Llama 3.2 1B, Llama 3.2 3B, Llama 3.1 8B, Gemma 3 1B, Gemma 3 4B

D Reward Model Training Dynamics

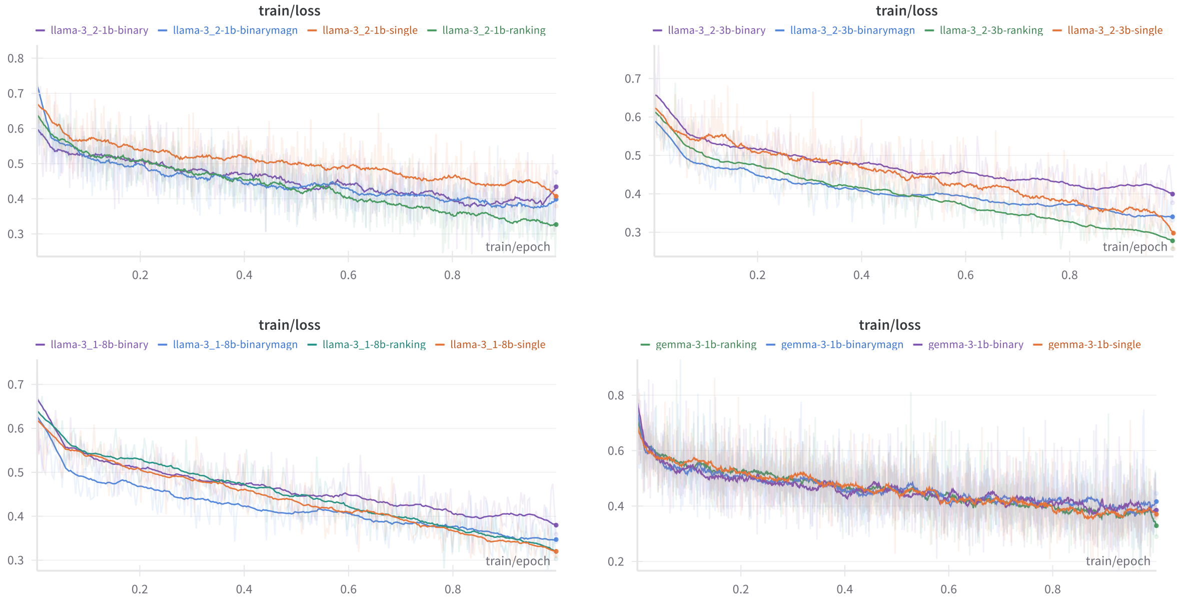

Figure 6: Reward Model Training dynamics