Modeling diverse preferences in movie artwork personalization with large language models

Hyunji Nam [1 2] Sejoon Oh [2] Emma Yanyang Kong [2] Yesu Feng [2] Moumita Bhattacharya [2]

Abstract

Large language models (LLMs) have demonstrated remarkable success in various recommendation applications. A key challenge within these systems is respecting diverse user preferences and value rather than offering a one-size-fits-all solution. Our work focuses on pluralistic preference alignment of LLMs for artwork recommendation on entertainment platforms, where systems need to consider diverse cultural norms, social values, and individual preferences when engaging with users. On these platforms, users typically interact with an extensive catalog of titles, each represented by specific artwork. Just as users' tastes are multi-faceted, titles contain varied themes and tones that may appeal to different viewers based on their values. Given this heterogeneity, we explore the novel problem of personalizing artwork recommendations using LLMs. For example, the same title might feature both heartfelt family drama and intense action scenes; a user preferring romantic content may favor artwork emphasizing emotional warmth, whereas a user preferring thrillers may find high-intensity scenes more appealing. We post-train 3B and 8B Llama3 models to select the optimal artwork for a given title-user pair based on the specific user's preferences. Our experiments with 110K training data and 5K held-out test data show that post-training yields a 3-5% improvement over a non-LLM production baseline. Overall, our work suggests a promising direction for hyper-personalized artwork recommendations, extending beyond text-based recommendation tasks and providing a pathway for pluralistic alignment with visual data.

[1]Stanford University (Work completed during the internship at Netflix) [2]Netflix. Correspondence to: Hyunji Nam <hjnam@stanford.edu>, Sejoon Oh <sejoono@netflix.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Large language models (LLMs) have demonstrated success in various applications of user recommendation and personalization across e-commerce and entertainment (Lyu et al., 2024; Lin et al., 2024; Yang et al., 2023; Dai et al., 2023; Lubos et al., 2024; Gao et al., 2023). Users are typically represented through natural language descriptions of their backgrounds, demographic information, or relevant interaction histories, and LLMs generate personalized predictions of user behavior or optimal actions that can improve user satisfaction or retention (Soni, 2023; Chen et al., 2024). Furthering this line of research, we present a novel task of personalized artwork recommendations based on individual users' tastes and preferences. On many existing entertainment platforms, users interact with a wide range of titles, each represented by a representative image (artwork). These artworks are often the first visual cues that influence users' decision to engage with titles, and therefore, have been explored as an important aspect of content personalization (Chandrashekar et al., 2017). As users exhibit diverse preferences, an artwork that appeals to one type of user may not resonate with another. Therefore, representing every title by a single artwork may fail to capture the varied interests of different users, especially across different age groups, regions, and cultural preferences.

We extend the relatively well-studied topic of text-based personalized recommendations using LLMs to the novel, yet crucial, sub-task in the visual domain: Modeling pluralistic user preferences among multiple visual representations of a given movie (title). We leverage the advanced capabilities of LLMs to process long input contexts, such as users' interaction histories, to decide which artworks would appeal the most to the users. While prior works have focused on title personalization (Yang et al., 2023; Wu et al., 2025) (for example, using the 1-million MovieLens dataset (Harper & Konstan, 2015) to predict users' ratings or make new title recommendations), we assume that the title is already selected, and focus on predicting which artwork from a set of candidate images would appeal most to the specific user based on their past interaction data. The number of artwork candidates may vary by titles, ranging from as few as four to more than forty, and the degree of visual differences across artworks also varies. While this keeps the realistic complexities of the problem, it potentially makes model training difficult, as the models need to learn to work with different-sized input contexts, varying candidate list sizes, and varying granularity in visual differences. Despite these challenges, our experiments with Llama 3.2-3B and 3.1-8B (Weerawardhena et al., 2025) show that post-training LLMs can achieve a performance improvement of 3–5%, compared to the non-LLM-based production model, on the held-out user-movie tuples. Specifically, we focus on two post-training methods: supervised fine-tuning with reasoning distillation from a stronger model (Qwen-32B (Team, 2025)) and Direct Preference Optimization (DPO) on chosen and rejected artwork pairs. These empirical results suggest a promising path for using LLMs in fine-grained, personalized content recommendations (e.g., artwork, synopsis, trailer) to better serve diverse user interests and preferences.

2. Related Work

LLMs for personalized recommendations Numerous works have explored using LLMs for personalizing content and product recommendations to overcome the limitations of conventional recommendation models (e.g., lack of world knowledge, limited text understanding, and inadequate reasoning capabilities) and have been successful at these tasks (Lyu et al., 2024; Lin et al., 2024). For example, Yang et al. (2023) leverage user-item interaction history expressed in natural language to personalize movie and product recommendations using LLMs. Dai et al. (2023) evaluate ChatGPT's inference-time capabilities for point-wise (i.e., predicting a user's rating for a single product), pair-wise (i.e., predicting a user's preference between an item pair), and list-wise (i.e., ranking a list of candidate items based on the user's previous examples) recommendations across movies, books, music, and news datasets without any additional training. Gao et al. (2023) explore the possibility of combining interactive LLM agents with traditional recommendation systems for cold-start and cross-domain generalization problems, as well as using LLMs to provide explanations for the recommendation system's personalization mechanism. Lubos et al. (2024) specifically focus on LLMs' abilities to generate explanations to improve transparency and engender user's trust in recommendation systems that have traditionally remained a black-box to the users. Unlike prior work focused on prompting and in-context learning from users' interaction histories, Wu et al. (2025) propose training a summarizer that can produce user-specific summaries with reinforcement learning from prediction feedback (RLPF) and using the generated summaries as user representations to personalize future product recommendations. Our work adds to this extensive line of research with a novel contribution. To our knowledge, we are the first to investigate the capabilities of LLMs to **personalize artwork selection among a set of image candidates for a given title**.

LLM post-training Post-training methods have been successfully applied to improve and refine the existing capabilities of LLMs trained on large-scale data (Kumar et al., 2025; Fernando et al., 2025). Popular strategies for post-training include supervised fine-tuning (SFT) and preference learning through reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022) or direct preference optimization (DPO) (Rafailov et al., 2024). Knowledge distillation has also been widely explored as a technique for training smaller models (e.g., open-source Llama models) to adopt the capabilities of larger models, including proprietary models like GPT-4 (Gu et al., 2025; Xu et al., 2024), where the training data is typically generated by the larger model and provided to the smaller model for fine-tuning. Other works have explored training and augmenting language models with reasoning abilities by having models explain their predictions for commonsense question-answering tasks (Rajani et al., 2019) or generating step-by-step solutions to math problems (Hendrycks et al., 2021) rather than directly outputting the final answers.

Automated image captioning There are various approaches to representing images through text captions (Li et al., 2023b; Mokady et al., 2021), with more recent work based on multimodal large language models (Meta AI, 2024) that can perform visual question answering (Guo et al., 2023), reasoning (Alayrac et al., 2022), and image captioning (Li et al., 2025; 2023a; Chen et al., 2023). While directly leveraging a visual language model is also an option, in this work we first convert the images into text captions via an image captioning model, and then use only the text inputs to train the model on personalized selection tasks.

3. Problem setup

3.1. Data composition

We have a dataset $\mathcal{D}$ consisting of users U, titles X, and multiple artwork options $A_{1:m}(X)$ for each title. Users are represented by their interaction histories which may include the user's viewing history and watch times, and the number of artwork options m varies across titles. For each user-title pair, we also have ground-truth optimal artwork A^* obtained from the user's previous engagement with the selected title. All artworks other than A^* are treated as negative examples that the user does not prefer. In our personalized recommendation setting, the optimal artwork depends on both the title and the user, $A^*(U, X)$.

For a test dataset $\mathcal{D}'$, we have a new set of user-title pairs. The goal of the recommendation system is to predict A^* among a set of candidates $A_{1:m}$ for a given (U, X) pair. We clarify that the training dataset may have overlapping users and titles, but no shared user-title pairs between train and test, so the recommendation system needs to generalize to unseen user-title combinations.

3.2. Evaluation metric

We consider two metrics: **accuracy** and **inverse propensity score (IPS)**. While accuracy is more intuitive to interpret and broadly applicable, IPS is better suited for our setting, where the number of artwork options m varies largely by titles.

Accuracy is defined as:

$$\sum_{i \in \mathcal{D}'} \frac{1}{|\mathcal{D}'|} I\{\hat{A}^*(u_i, x_i) = a_i^*\}, \tag{1}$$

where $\hat{A}^*$ is the model's prediction for the optimal personalized artwork for a given tuple (u_i, x_i). One limitation with accuracy is that it does not distinguish between cases with 40 artwork options versus 2 options. Intuitively, selecting the optimal artwork from a set of two has a higher chance of being correct by random selection compared to selecting the optimal artwork from a larger set. Therefore, a more nuanced evaluation metric would account for the number of artworks available for each title and appropriately reflect the difficulty of a correct selection by chance.

This leads to **IPS**, which is defined as:

$$\sum_{i \in \mathcal{D}'} \frac{1}{|\mathcal{D}'|} \frac{I\{\hat{A}^*(u_i, x_i) = a_i^*\}}{\pi(a_i^*)}. \tag{2}$$

$\pi(a_i^*)$ denotes the probability of the ground truth artwork a_i^* being shown to the user. By setting π to be uniform over the set of artworks, we now have a metric that accounts for the difficulty of selecting the correct artwork from sets of different sizes. For example, a correct prediction from a set of 40 artwork candidates is weighted twenty times higher than a correct prediction from a set of two. Due to the varying number of artwork options across different films in the catalog, IPS is the primary evaluation metric used by the production system.

4. Method

4.1. Task formulation

For the user representation X, we use the user's most recent K timestamped interactions, which may include information such as the user's recently engaged titles and their genres. Each artwork is represented by a caption (of $\approx$ 200 tokens) generated by a fine-tuned Llama-3.2-11B visual language model (Meta AI, 2024). We formulate the task as a prediction problem: given the user's past interactions, a new title, and a list of artworks each represented by a caption, predict which artwork would most appeal to the user. Representing the images as captions enables all multimodal inputs to be expressed in text, which is easily supported by existing LLM post-training methods, and the compactness of the text captions also allows our method to scale to a larger number of candidate artworks (e.g., 40+ options for certain titles).

We conduct post-training of Llama 3.1 8B models (Weerawardhena et al., 2025). The prompt describes the user's interaction history and lists a set of artwork options for a new title, and the model's objective is to predict which artwork the user is most likely to prefer based on their taste and interest.

We introduce two new tokens, `<option>` and `</option>`, to delimit different artwork options in the prompt. During inference, we guide the model's generation with the prefix "Prediction: `<option>`." Additionally, to extract the predicted caption, we apply n-gram matching to the model's output to select the artwork option with the highest match as the model's choice.

4.2. Post-training methods

- **Supervised fine-tuning (SFT).** We use (u_i, x_i, a_i^*) tuples from the training dataset (n=110K) to fine-tune Llama 3.2 3B-Instruct (Meta, 2025) and Llama 3.1 8B-Instruct (Weerawardhena et al., 2025) using the prompt and expected model output shown above. The SFT loss is defined as:

$$\mathcal{L}_{\text{SFT}}(\pi_\theta) = -\mathbb{E}_{i \sim \mathcal{D}} \Big[\log \pi_\theta(a_i^* | x_i, u_i) \Big] \tag{3}$$

- **Prediction with reasoning.** We leverage the recently advanced reasoning capabilities of LLMs to enable personalized artwork selection. Specifically, we augment the training data with a powerful reasoning model, Qwen/QwQ-32B (Team, 2025), to generate explanations conditioned on the known ground truth a_i^*. Alternatively, one could first generate reasoning along with predictions, filter out examples that do not match the correct artwork, and repeat this reasoning generation process to obtain a high-quality reasoning dataset (Zelikman et al., 2022). However, due to the difficulty of our prediction task, off-the-shelf Qwen models achieve low prediction accuracy, which would lead to extensive re-sampling to obtain reasoning outputs that are consistent with the ground-truth answers, and is therefore, difficult to scale with the larger training data. In order to avoid the expensive re-sampling process, we first provide the reasoning model with the ground-truth artwork selection, and prompt it to generate a matching explanation. The model also predicts the most appropriate artwork to recommend conditioned on its own explanation. We additionally filter out the model's responses that still do not match the ground-truth labels (which accounts for less than 2% of the model's generations). Despite the model's initial low prediction accuracy, this approach allows us to generate justifications that support the ground truth, which can then be used to construct the reasoning dataset for training. We augment the training data with the generated reasonings, so each tuple (u_i, x_i, a_i^*) is additionally tagged with the corresponding reasoning r_i. We then supervised fine-tune 8B Llama-Instruct (Weerawardhena et al., 2025) using this reasoning-augmented dataset, with the expected model output starting with "Reason: ..." followed by the item prediction.

- **Direct policy optimization (DPO).** We train a 8B Llama model (Weerawardhena et al., 2025) using the DPO objective (Rafailov et al., 2024), where the chosen-rejected response pair for each tuple (x_i, u_i) consists of a_i^* (chosen) and a randomly selected alternative a_i' (rejected) from the remaining candidate set. The DPO loss objective[1] (Rafailov et al., 2024), $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$

$$= -\mathbb{E}_{i\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(a_i^* \mid x_i, u_i)}{\pi_{\text{ref}}(a_i^* \mid x_i, u_i)} - \beta\log\frac{\pi_\theta(a_i' \mid x_i, u_i)}{\pi_{\text{ref}}(a_i' \mid x_i, u_i)}\right)\right] \tag{4}$$

leads the recommendation model to up-weight the likelihood of a preferred artwork and down-weight the likelihood of a rejected artwork. This can be more effective than training with only positive examples, as the model explicitly learns to distinguish between positive and negative responses. On the other hand, SFT-only model may learn the instruction-following behavior but does not get to observe how negative examples compare to positive examples despite the structural similarities in their response patterns. DPO can be combined with SFT by first training the base model using the SFT objective and then continuing training with the DPO objective, as typically done in other works (Chen et al., 2025; Liu et al., 2024).

[1] β is a hyperparameter that determines how close the trained model should be to the base model, π_{ref}, for maintaining stability.

Table 1. We conducted experiments with Llama 8B on a training dataset of 110K examples and evaluated on 5K held-out user-title tuples. *The reported values are percentage improvements compared to the production model*, which is a non-LLM deep neural network trained on a large-scale production dataset.

Method	Accuracy	IPS
Random guess	-74.96%	-4.59%
Zero-shot prediction	-4.22%	-0.19%
Baseline	0%	0%
+ SFT	-2.55%	+2.45%
+ DPO	**+0.91%**	**+2.82%**
+ SFT with reasoning from Qwen-32B	**+1.41%**	**+5.21%**

Personalized recommendation model instruction

Sample prompt template: "You are an expert in movies and shows. I want you to predict which of the available artworks the user would like the most based on their past watch history. [User's past interaction data] The user's new title is: [new title]. Here are the artwork options: <option> caption describing A_1 </option>, <option> caption describing A_2 </option>, ... ,<option> caption describing A_m </option>. Output the best artwork in text.
Model output: "Prediction: </option> caption describing $\hat{A}^*$ </option>."

All models are trained using low-rank adaptation (LoRA) (Hu et al., 2021). We conduct a hyperparameter search across different learning rates (e.g., $\{1e-7, 5e-7, 1e-6, 5e-6, 1e-5, 1e-4\}$) and report evaluation results obtained with the best-performing model from the validation set of size 1,000. The evaluation dataset has 5K new user-title pairs.

5. Experiment & Results

5.1. How does post-training LLMs perform compared to the production model?

Table 1 shows that the post-trained models achieve a performance improvement of 3-5% in terms of IPS compared to the production model. Surprisingly, even the zero-shot model achieves decent performance significantly better than random guessing. This speaks to the power of world knowledge and long-context processing capabilities that the current large language models are equipped with that can lead the models to make reasonable predictions about user behavior even though they are not trained on any domain-specific (containing information about users or titles) dataset.

Table 2. We compare the zero-shot performance of the Llama 3 and 8B models on a test set of 5K user–title pairs *to the production model's performance in terms of percentage difference.* The smaller the difference, the closer the zero-shot model's performance matches the production level.

Model size	Output format	Accuracy	IPS
3B	**Number**	**-11.55%**	**-6.81%**
3B	Text	-12.76%	-8.15%
8B	Number	-7.06%	-1.73%
8B	**Text**	**-6.52%**	**-0.19%**

5.2. How sensitive is the model to number versus text based prediction?

Many prior works have observed the large language model's sensitivity to different prompting perturbations, like instruction templates and paraphrasings (Berglund et al., 2024; Mizrahi et al., 2024; Zhuo et al., 2024) . One particular dimension of sensitivity we investigate is the output format: whether the model is instructed to format its prediction as an integer representing the artwork from the given list, or as a full-text caption describing the selected artwork. In order to decide on the output format, we conduct an ablation with zero-shot 3B and 8B Llama models comparing their performance on text output versus artwork option number output. Surprisingly, we observe that with the 3B models, the number output outperforms the text output in terms of the evaluation metrics, which is different from the performance of the 8B models (Table 2). However, a breakdown analysis of the model's accuracy across different ground truth labels (Fig. 1) reveals the model's tendency to output smaller numbers. In particular, the 3B model prompted to predict the artwork option number shows 0% accuracy for examples with ground truth labels larger than 15, suggesting that the model's high accuracy is compounded by its bias toward predicting smaller numbers.

5.3. How does post-training impact different base models?

In the previous section, we showed that SFT with reasoning, where the models are trained to output the chosen artwork in text, achieves the highest post-trained performance. Ablations in Table 3 further test whether this trend holds for different model sizes and training dataset sizes. Specifically, these ablations show that training with a reasoning dataset and outputting text predictions achieves the highest performance gains for both 3B and 8B base models, and for both 10K- and 110K-sized training datasets. As expected, the performance gains from the larger dataset are greater than those from the smaller dataset for the same model size. Surprisingly, we observe that the smaller models obtain larger

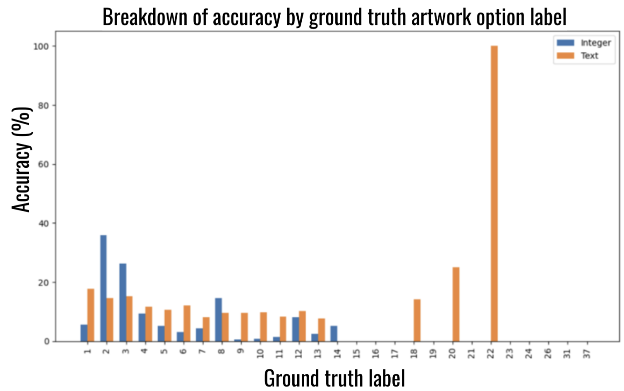

Figure 1. **A breakdown of accuracy across different ground truth labels to compare the performance of the models outputting the artwork option number versus the artwork caption in full text.** The x-axis shows the ground truth label (artwork option number) and the y-axis shows the model's prediction accuracy for samples with a particular ground truth label. Although the average performance across all examples is higher for the number prediction (blue) compared to the text caption prediction (orange), the breakdown by the ground truth label suggests that number prediction performs poorly for higher-valued ground-truth.

Table 3. **How much delta improvement is made relative to the zero-shot 3B & 8B base model?** We conducted experiments with Llama 3 & 8B using LoRA-SFT and evaluated with 5K data points. We report the percentage improvements *after post-training compared to the base model of the same type.*

Model	Output	Training size	Accuracy △	IPS
3B	Zero-shot	NA	0%	0%
+ 3B	Number	10K	6.71%	5.76%
+ 3B	Text	10K	8.38%	8.06%
+ 3B	**Text + Reason**	**10K**	**10.53%**	**10.60%**
8B	Zero-shot	NA	0%	0%
+ 8B	Number	10K	-0.40%	0.67%
+8B	Text	10K	1.14%	1.52%
+ 8B	**Text + Reason**	**10K**	**3.05%**	**3.02%**
+ 8B	Number	110K	1.64%	-0.96%
+ 8B	Text	110K	1.67%	3.02%
+ 8B	**Text + Reason**	**110K**	**5.63%**	**5.40%**

performance improvements from post-training. However, the final performance of the 8B models still dominates that of the 3B models when compared to the production model's performance.

Table 4 compares the DPO-trained performance of different checkpoints. While we observe that DPO from the SFT model checkpoint outperforms DPO from the off-the-shelf pretrained model, it does not outperform SFT with reasoning alone. We suspect that the negative examples used to construct the preference pairs may not be clearly distinguished from their positive counterparts. For example, the current training dataset has a single ground-truth optimal artwork which the user has engaged with and liked, while the remaining artworks for the same title (i.e., counterfactuals) are assumed to be negative. However, it is unclear

Table 4. We compared the performance of DPO from various checkpoints. All the experiments were conducted with Llama-8B. *The reported values are percentage improvements compared to the production model.*

Training size	SFT checkpoint?	Accuracy	IPS
10K	(SFT without reasoning)	-4.26%	+1.42%
10K	(SFT with reasoning)	-3.73%	0%
110K	(from pre-trained model)	-5.38%	-0.19%
110K	**(SFT with reasoning)**	**-3.21%**	**+2.82%**

whether a negative example was skipped by the randomized artwork selection algorithm in production, and therefore was not shown to the user.

6. Discussion & Limitation

In summary, our work introduces the novel recommendation task of predicting personalized artwork for different user-title pairs. Specifically, we provide LLMs with a verbalized representation of a user's interaction history, which reveals their content preferences, and ask the model to predict which artwork caption from a candidate set is most likely to appeal to the given user. We leverage the LLM's natural language reasoning capabilities and show that the model's world knowledge augmented with post-training on 110K user data achieves a 5% performance improvement compared to the production model. Additionally, in our hyperparameter search, we observe that smaller learning rates generally achieve better training performance, as similarly observed by Pareja et al. (2024). However, constructing a reliable user dataset of chosen and rejected in the presence of missing counterfactuals remains to be a challenge.

Our work suggests the following promising directions for personalized recommendation systems: **(1) Improving models with reasoning.** Based on the improvements we observed with SFT with reasoning compared to vanilla SFT, we expect DPO with reasoning to further improve the model's performance and can also lead to transparent and interpretable AI systems that can provide explanations for the model's decisions. **(2) Leveraging multi-modal LLMs.** While our work represents artworks with captions, recent advances in multimodal LLMs (OpenAI, 2025; Google Deepmind, 2025) suggest that the images can be directly embedded into the instruction to avoid information loss from image to text transformation. **(3) Personalizing various aspects of content recommendation.** We encourage future work to consider extending this framework to other components of content personalization beyond artworks (e.g., synopses, trailers), or even to directly recommend new artwork designs in natural language, which service providers and artists can use for creative generation to satisfy diverse user preferences and values. Overall, our work suggests a promising direction for hyper-personalized artwork recommendations, extending the utility of LLMs beyond traditional text-based recommendation and providing a pathway for pluralistic alignment with visual input data.

Impact Statement

This work focuses on the personalization of artworks that can respect diverse, heterogeneous user preferences. Incorporating LLMs into recommendation systems opens up a path to improving the accuracy of predicting users' preferred content, as well as enhancing the transparency of the recommendation models by generating interpretable natural language explanations of the model's selections. However, any application that requires handling user data must be implemented with care and monitored to ensure compliance with user privacy policies.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2024. URL https://arxiv.org/abs/2309.12288.

Chandrashekar, A., Amat, F., Basilico, J., and Jebara, T. Artwork personalization at netflix, Dec 7 2017. URL https://netflixtechblog.com/artwork-personalization-c589f074ad76. Accessed: 2025-11-25.

Chen, C., Liu, Z., Du, C., Pang, T., Liu, Q., Sinha, A., Varakantham, P., and Lin, M. Bootstrapping language models with dpo implicit rewards, 2025. URL https://arxiv.org/abs/2406.09760.

Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Zheng, K., Lian, D., and Chen, E. When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web*, 27(4), June 2024. ISSN 1573-1413. doi: 10.1007/s11280-024-01276-1. URL http://dx.doi.org/10.1007/s11280-024-01276-1.

Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., and Liu, Z. Large language models are visual reasoning coordinators, 2023. URL https://arxiv.org/abs/2310.15166.

Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., Sun, Z., Zhang, X., and Xu, J. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pp. 1126–1132. ACM, September 2023. doi: 10.1145/3604915.3610646. URL http://dx.doi.org/10.1145/3604915.3610646.

Fernando, H., Shen, H., Ram, P., Zhou, Y., Samulowitz, H., Baracaldo, N., and Chen, T. Understanding forgetting in llm supervised fine-tuning and preference learning – a convex optimization perspective, 2025. URL https://arxiv.org/abs/2410.15483.

Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., and Zhang, J. Chat-rec: Towards interactive and explainable llms-augmented recommender system, 2023. URL https://arxiv.org/abs/2303.14524.

Google Deepmind. Gemini 2.5: Our most intelligent ai model, Mar 25 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking. Accessed: 2025-11-25.

Gu, Y., Dong, L., Wei, F., and Huang, M. Minillm: Knowledge distillation of large language models, 2025. URL https://arxiv.org/abs/2306.08543.

Guo, J., Li, J., Li, D., Tiong, A. M. H., Li, B., Tao, D., and Hoi, S. C. H. From images to textual prompts: Zero-shot vqa with frozen large language models, 2023. URL https://arxiv.org/abs/2212.10846.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Kumar, K., Ashraf, T., Thawakar, O., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., Torr, P. H. S., Khan, F. S., and Khan, S. Llm post-training: A deep dive into reasoning large language models, 2025. URL https://arxiv.org/abs/2502.21321.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. Mimic-it: Multi-modal in-context instruction tuning, 2023a. URL https://arxiv.org/abs/2306.05425.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Cahyono, J. A., Yang, J., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning, 2025. URL https://arxiv.org/abs/2305.03726.

Li, W., Zhu, L., Wen, L., and Yang, Y. Decap: Decoding clip latents for zero-shot captioning via text-only training, 2023b. URL https://arxiv.org/abs/2303.03032.

Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Zhang, H., Liu, Y., Wu, C., Li, X., Zhu, C., Guo, H., Yu, Y., Tang, R., and Zhang, W. How can recommender systems benefit from large language models: A survey, 2024. URL https://arxiv.org/abs/2306.05817.

Liu, Y., Liu, P., and Cohan, A. Understanding reference policies in direct preference optimization, 2024. URL https://arxiv.org/abs/2407.13709.

Lubos, S., Tran, T. N. T., Felfernig, A., Polat Erdeniz, S., and Le, V.-M. Llm-generated explanations for recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, pp. 276–285, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704666. doi: 10.1145/3631700.3665185. URL https://doi.org/10.1145/3631700.3665185.

Lyu, H., Jiang, S., Zeng, H., Xia, Y., Wang, Q., Zhang, S., Chen, R., Leung, C., Tang, J., and Luo, J. LLM-rec: Personalized recommendation via prompting large language models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 583–612, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.39. URL https://aclanthology.org/2024.findings-naacl.39/.

Meta, A. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, september 2024. *URL https://ai. meta. com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed*, pp. 05–10, 2025.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models for mobile & edge devices, Sep 25 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices Accessed: 2025-11-25.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 08 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00681. URL https://doi.org/10.1162/tacl_a_00681.

Mokady, R., Hertz, A., and Bermano, A. H. Clipcap: Clip prefix for image captioning, 2021. URL https://arxiv.org/abs/2111.09734.

OpenAI. Introducing gpt-5, Aug 5 2025. URL https://openai.com/index/introducing-gpt-5/. Accessed: 2025-11-25.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Pareja, A., Nayak, N. S., Wang, H., Killamsetty, K., Sudalairaj, S., Zhao, W., Han, S., Bhandwaldar, A., Xu, G., Xu, K., Han, L., Inglis, L., and Srivastava, A. Unveiling the secret recipe: A guide for supervised fine-tuning small llms, 2024. URL https://arxiv.org/abs/2412.13337.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. Explain yourself! leveraging language models for commonsense reasoning. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL https://aclanthology.org/P19-1487/.

Soni, V. Large language models for enhancing customer lifecycle management. *Journal of Empirical Social Science Studies*, 7(1):67–89, 2023.

Team, Q. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Weerawardhena, S., Kassianik, P., Nelson, B., Saglam, B., Vellore, A., Priyanshu, A., Vijay, S., Aufiero, M., Goldblatt, A., Burch, F., Li, E., He, J., Kedia, D., Oshiba, K., Yang, Z., Singer, Y., and Karbasi, A. Llama-3.1-foundationai-securityllm-8b-instruct technical report, 2025. URL https://arxiv.org/abs/2508.01059.

Wu, J., Ning, L., Liu, L., Lee, H., Wu, N., Wang, C., Prakash, S., O'Banion, S., Green, B., and Xie, J. Rlpf: Reinforcement learning from prediction feedback for user summarization with llms, 2025. URL https://arxiv.org/abs/2409.04421.

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., and Zhou, T. A survey on knowledge distillation of large language models, 2024. URL https://arxiv.org/abs/2402.13116.

Yang, F., Chen, Z., Jiang, Z., Cho, E., Huang, X., and Lu, Y. Palr: Personalization aware llms for recommendation, 2023. URL https://arxiv.org/abs/2305.07622.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. D. Star: Bootstrapping reasoning with reasoning, 2022. URL https://arxiv.org/abs/2203.14465.

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., and Chen, K. Prosa: Assessing and understanding the prompt sensitivity of llms, 2024. URL https://arxiv.org/abs/2410.12405.