FACTUAL AND PERSONALIZED RECOMMENDATIONS USING LANGUAGE MODELS AND REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recommender systems (RSs) play a central role in connecting users to content, products, and services, matching candidate items to users based on their preferences. While traditional RSs rely on implicit user feedback signals, conversational RSs interact with users in natural language. In this work, we develop a *comPelling*, *Precise*, *Personalized*, *Preference-relevant* language model (P^{4}LM) that recommends items to users while putting emphasis on explaining item characteristics and their relevance. P^{4}LM uses the *embedding space* representation of a user's preferences to generate compelling responses that are factually-grounded and relevant w.r.t. the user's preferences. Moreover, we develop a joint reward function that measures precision, appeal, and personalization, which we use as AI-based feedback in a reinforcement learning-based language model framework. Using the MovieLens 25M dataset, we demonstrate that P^{4}LM delivers compelling, personalized movie narratives to users.

1 INTRODUCTION

Recommender systems (RSs) have emerged as a dominant way in which users discover content, products, and services (Resnick & Varian, 1997). Traditional RSs match candidate items to users based on their estimates for items preferences, possibly conditioned on some query or context. However, these preferences are often based on implicit user behavioral signals, such as clicks, number of watches, ratings, purchases, etc. Unfortunately, these provide little opportunity for an RS to elicit high-bandwidth preference information from users, explain recommendations, or for users to critique and steer their interaction with the RS. *Conversational RSs* have therefore attracted considerable attention as means to use natural-language interaction to facilitate more effective communication between RSs and their users (Sun & Zhang, 2018; Lei et al., 2020; Shen et al., 2023).

The emergence of language models (LMs) as a powerful paradigm for user engagement (Li et al., 2018; Friedman et al., 2023) suggests their use as a vehicle for conversational RSs. However, this requires LMs to engage in a personalized manner, adhering to users' preferences. In this paper, we explore the intersection of RSs and LMs, and more particularly, the use of LMs to enrich the user experience in RSs. We develop techniques which allow an LM to communicate the nuances of recommended items to a user, detailing their features, benefits, and explaining their *alignment with a user's preferences*. Such *personalized LMs* are not meant to "convince" users in the traditional sense, but rather, to articulate the *genuine and relevant merits* of a recommended item relative to the user.

Personalized LMs offer users a fully tailored RS experience, ensuring they find what they truly need and value. However, a number of challenges must be addressed in this endeavor: (i) any recommended item should be predicted to have maximal value given the user's preferences; (ii) the integrity and accuracy of an item's information is paramount; (iii) the personalized LM should present a reasonably comprehensive portrayal of the item by describing its merits and drawbacks, with a focus on *relevance* to the user's preferences; (iv) and finally, the LM's explanations or endorsements should be compelling and appealing to the user, provided that it meets the other criteria. In this work, we develop a framework centered around these four principles.

A key question we addressed in this work is how to effectively utilize the information captured by an RS embedding space to generate a factual, personalized, compelling, and relevant recommendations.

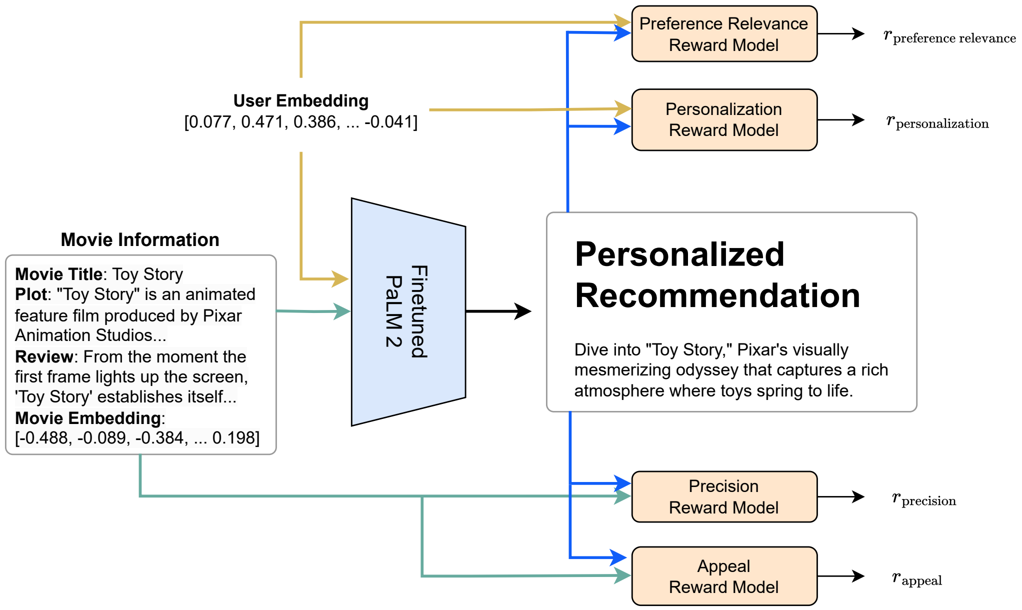
User Embedding
[0.077, 0.471, 0.386, ... -0.041]
Movie Information
Movie Title: Toy Story
Plot: "Toy Story" is an animated feature film produced by Pixar Animation Studios...
Review: From the moment the first frame lights up the screen, 'Toy Story' establishes itself...
Movie Embedding:
[-0.488, -0.089, -0.384, ... 0.198]
Finetuned PaLM 2
Personalized Recommendation
Dive into "Toy Story," Pixar's visually mesmerizing odyssey that captures a rich atmosphere where toys spring to life.
Preference Relevance Reward Model
Personalization Reward Model
Precision Reward Model
Appeal Reward Model
$r_{\text{preference relevance}}$
$r_{\text{personalization}}$
$r_{\text{precision}}$
r_{appeal}

Figure 1: The P^{4}LM Learning Framework for Recommendation Endorsement Generations.

Our contributions are three-fold. First, we quantify the aforementioned four attributes using reward functions, enabling systematic evaluation. Second, leveraging recent advances in reinforcement learning from AI feedback (RLAIF) (Lee et al., 2023), we develop an LM fine-tuning methodology to better align with these four rewards (see Figure 1 for the schematic diagram illustrating the RLAIF framework). Our developed model, which we term P^{4}LM, not only comprises semantic skills, but also understands users' preferences encoded in the RS embedding space, providing factual, compelling, personalized endorsements. Finally, building on the MovieLens 25M dataset (Harper & Konstan, 2015) we showcase the potential of P^{4}LM , powering a conversational movie recommender that promotes customized, relevant, and holistic interactions for users.

We begin with a brief introduction of RSs, LMs and the use of contextual Markov decision processes (CoMDPs) for modeling generative language problems of RSs (Section 2). We then describe the four principles, (i.e., personalization, precision, appeal, and preference relevance), which we incorporate into training of LMs for RSs (Section 3), followed by an reinforcement learning based fine-tuning methodology for training P^{4}LM (Sections 4). Finally, we demonstrate the effectiveness of P^{4}LM in generating factual, personalized, and compelling movie endorsement narratives for users within the MovieLens 25M benchmark dataset (Section 5).

2 Preliminaries

In this section we present some basic background, outline our problem formulation, and establish the terminology used throughout the paper.

Recommender Systems (RSs). To model user-item behavioral relationships in a personalized RS, we assume a standard collaborative filtering (CF) task (Su & Khoshgoftaar, 2009). Collaborative filtering finds similar patterns among users, filtering out items based on ratings of similar users. Given a user $u \in \mathcal{U}$, we use $r_{u,i}$ (e.g., 1–5 stars) to denote the rating of item $i \in \mathcal{I}$ by user u. Let $\mathcal{R}$ denote the $|\mathcal{I}| \times |\mathcal{U}|$ (usually sparse) ratings matrix corresponding to the *ratings dataset* $\mathcal{R} = \{(u, i, r_{u,i}) : r_{u,i} \neq 0\}$. To predict users' preference behavior, an RS learns user and item representations from the ratings dataset $\mathcal{R}$ using a CF approach. Then, the resulting *item embedding* maps each item i to a vector representation $\mathbf{i}$ of its (latent) attributes. Note that these embeddings are typically not interpretable. Similarly, user preferences are captured by a *user embedding*, mapping users u to a vector representation $\mathbf{u}$.

Methods including matrix factorization (Mnih & Salakhutdinov, 2007) or neural CF (Rendle et al., 2020; He et al., 2017; Beutel et al., 2018) are used to learn the user and item embeddings, which assumes a *two-tower model* (or *dual encoder*) in which users and items are passed through separate (but co-trained) deep neural nets (DNNs) to produce their respective vector embeddings $\mathbf{u}$ and $\mathbf{i}$. These are then combined via dot product to predict user-item affinity $\hat{r}_{i,u}$ (Yi et al., 2019; Yang et al., 2020). We view $\mathbf{i}$ as a (learned) latent feature vector characterizing item i and $\mathbf{u}$ as parameterizing user u's estimated *utility (or preference) function* over these features.

Language Models (LMs). In this work, we inject a user's behavioral information into a seq2seq LM (Vaswani et al., 2017) to generate personalized recommendation responses. We assume a dataset of the form $\mathcal{D} = \{(\mathbf{I}^{(k)}, \mathbf{i}^{(k)}, \mathbf{u}^{(k)}, Y^{(k)})\}_{k=1}^{|\mathcal{D}|}$, where $\mathbf{I}$ is a textual description of some item $i \in \mathcal{I}$ (e.g., descriptions, positive/negative reviews from different users); $\mathbf{i}$ is the CF embedding vector of i; $\mathbf{u}$ is the CF embedding vector of a user $u \in \mathcal{U}$; and finally, Y is a textual response (e.g., compelling recommendation, endorsement or explanation) tailored to the user. We refer to Appendix C for details on the generation of $\mathcal{D}$.

Let $N_{\mathbf{I}}$ be an upper-bound on the length (number of tokens) of any item description $\mathbf{I}$.[1] The role of an LM is to predict the probability $\mathbb{P}\big(Y = \{y_n\}_{n=0}^{N-1} \mid y_0, \mathbf{I}, \mathbf{i}, \mathbf{u}\big)$ of the personalized response Y (N tokens), conditioned on the item description $(\mathbf{I}, \mathbf{i})$ and user embedding $\mathbf{u}$.

In standard LMs, a Transformer (Wolf et al., 2019) architecture T encodes an item's textual context $\mathbf{I}$ as an $N_{\mathbf{I}}$-length sequence of embeddings $(z_0, \ldots, z_{N_{\mathbf{I}}-1})$ induced by the transformer's attention layers. For convenience, we concatenate these into a single embedding $z \in \mathcal{Z} \subseteq \mathbb{R}^d$, where d is the dimension of the latent space. The text response $Y = \{y_n\}_{n=0}^{N-1}$ is sampled token-by-token in an auto-regressive manner using a decoder Ψ; i.e., $Y \sim \Psi\big(\cdot \mid z\big) := \prod_{n=0}^{N-1} \Psi\big(y_n \mid y_0, \ldots, y_{n-1}; z\big)$, where y_0 is a fixed start-of-sentence token (Chien & Kuo, 2019). To incorporate behavioral information into the LM, the standard LM is augmented with adapters (Pfeiffer et al., 2020) $\mathrm{W}_I, \mathrm{W}_U : \mathcal{V} \mapsto \mathcal{Z}$, to induce the language model: $\Psi \circ (\mathrm{T} \times \mathrm{W}_I \times W_U)$ (Jaech & Ostendorf, 2018). Here, T maps text-input tokens to $\mathcal{Z}$ whereas W_I (resp., W_U) maps item (resp., user) CF-embedding vectors $\mathcal{V}$ to $\mathcal{Z}$. Importantly, T, W_I, and W_U map tokens and CF vectors to a common space so that their relationship can be captured by the transformer's attention mechanism.

Contextual Markov Decision Processes (CoMDPs). CoMDPs have been used to model token-wise generative language problems (Li et al., 2016; Asadi & Williams, 2016; Jaques et al., 2019), and can also be used in conversational RSs. In this MDP, the LM acts as a policy which maps text inputs and user/item behavioral embedding vectors to generated responses.

Let $(\mathcal{C}, \mathcal{S}, \mathcal{A}, P, r, s_0, N)$ denote the CoMDP, where the observable context space $\mathcal{C}$ contains item/user information $\mathcal{I}$, $\mathbf{i}$ and $\mathbf{u}$. The horizon N is the length of the generated text. The state space $\mathcal{S}$ at the n-th turn ($n < N$) is the sequence of tokens $\{y_0, \ldots, y_{n-1}\}$ generated thus far, with s_0 being the start-of-sentence token y_0. The action space $\mathcal{A}$ is the language token vocabulary, with action $a \in \mathcal{A}$ representing any possible next token. The transition kernel P models the next token distribution given the current sequence and contexts, which coincides with the LM policy (and is thus known). Finally, the reward function r measures the overall quality of the generated text. Our goal is to find a policy π^* which achieves maximum expected cumulative return, i.e., $\pi^* \in \arg\max_\pi J_\pi := \mathbb{E}[\sum_{n=0}^{N-1} r_t \mid P, s_0, \mathcal{C}, \pi]$. Note that the size of the tokenized state and action spaces grow exponentially with the vocabulary size.

3 Factual & Personalized Recommendations with LMs

A key question when using LMs for recommendation is how to effectively use the information captured by the RS embedding space to generate a factual, personalized, compelling, and relevant text response. Treating an LM as a factored distribution of item-user information over generated text tokens, one standard approach is to learn this model with *behavioral cloning (BC)* (Sasaki &

[1] If the actual description $\mathbf{I}$ has fewer tokens than $N_{\mathbf{I}}$, remaining spaces in the utterance will be padded by a specific token and masked.

Yamashina, 2020), by maximizing the conditional log-likelihood w.r.t. to the dataset $\mathcal{D}$:

$$\min_{\Psi} \ L_{\text{Cond}}(\Psi) := -\mathbb{E}_{(\mathbf{I},\mathbf{i},\mathbf{u},Y)\sim D}[\sum_{n=0}^{N-1} \log \Psi(y_n \mid y_0, \ldots, y_{n-1}; \mathbf{I}, \mathbf{i}, \mathbf{u})].$$

While this model may learn to interpret the behavioral information captured in the RS embeddings, the LM might actually lean towards disregarding the embedding contexts due to the typically more predictable nature of token generation when given text inputs. Consequently, the model might concentrate solely on text information, effectively degenerating to a non-contextual LM. To prevent this from occurring, and more importantly to ensure the LM can offer a comprehensive RS experience, we incorporate four key metrics into our training procedure; namely, *personalization*, *precision*, *appeal*, and *preference relevance*. We detail these next.

Precision. LM-based personalized recommendation can be viewed as a special form of abstractive summarization (Zhang et al., 2020a; Liu et al., 2022): the generated text should capture item characteristics that explain why a user would benefit from the recommendation. To preserve the RS's integrity, of course, one must emphasize *truthfulness* in its recommendation. That is, the RS's generated recommendation should describes genuine merits (and drawbacks) of the item, rather than persuasive distortions.

While recent summarization techniques produce highly coherent texts, they often suffer from *hallucinations* (Ji et al., 2023) – the tendency to generate information unsupported by the input text. Such factual inconsistencies may therefore limit their real-world applicability. Inspired by Roit et al. (2023) and Honovich et al. (2022), we evaluate factuality in our LM-based RS using an *entailment reward* (Bowman et al., 2015). Unlike widely-used metrics, such as ROUGE (Lin, 2004), that are ineffective at hallucination detection, we adopt a *textual entailment* (or natural language inference (NLI)) metric to measure truthfulness of our generated text, viewing it as a partial summary of an items's description. Particularly, given a description $\mathbf{I}$, we define the NLI score $\text{NLI}(Y; \mathbf{I})$ of text-token sequence Y as the probability of entailment under a classifier trained on several textual entailment datasets (see e.g., MacCartney & Manning (2007)). While this metric is not specifically tailored to summarization tasks, Honovich et al. (2021) show that it effectively detects factual inconsistencies in generated text. Since faithful summaries should be textually entailed by the input documents, such a metric provides informative feedback about the precision of generated item texts.

Of course, factual entailment is clearly insufficient in and of itself. In fact, it is rather easy to optimize a degenerate response which maximizes factual entailment (e.g., producing summaries that are highly extractive (Ladhak et al., 2021) or uninformative (Skalse et al., 2022)). In what follows we describe three other metrics we require for a comprehensive recommendation experience.

Appeal. Recent work has paid increasing attention to enriching recommendations to appeal to users (Felfernig et al., 2007; Zhang et al., 2020b). To the extent that we do not sacrifice user welfare, personalization, or factuality, such recommendations have value as they encourage users to accept recommendations of high personal utility. With recent LM technologies (Google et al., 2023; OpenAI, 2023), a plausible approach is to simply prompt an LM to generate an *endorsement* to complement its item recommendation. Such an endorsement, apart from being factual, should be compelling for the user. However, without systematic evaluation of such methods (e.g., do users find them appealing or compelling), it remains unclear whether they can improve the user experience. Quantifying appeal is challenging, as it may depend on subjective factors such as *style* (concise phrases over detailed explanations) and *language* (compelling, eloquent pitches over dry factual summaries).

To assess appeal, we use a dataset of pairwise human/machine demonstrations (see Appendix C for details on its construction). We develop an *appeal model* which scores the generated text Y and assess how compelling it are, using learning from human/AI feedback (LAIF) (Christiano et al., 2017). Specifically, let $\mathcal{D}_{\text{app}} = \{(Y_w^{(k)}, Y_l^{(k)}; \mathbf{I})\}_{k=1}^{|\mathcal{D}_{\text{app}}|}$ be a labeled dataset reflecting the relative appeal of two recommendation texts Y_w, Y_l given textual item description $\mathbf{I}$. Here, $Y_w \succ Y_l | \mathbf{I}$ indicates that Y_w is more compelling given $\mathbf{I}$. Assuming these relationships are governed by a latent model $\text{App}(Y; \mathbf{I})$, we parameterize it via Bradley-Terry (Huang et al., 2006), where the appeal distribution is defined by

$$p_{\text{app}}(Y_w \succ Y_l; \mathbf{I}) = \frac{\exp(\text{App}(Y_l; \mathbf{I}))}{\exp(\text{App}(Y_w; \mathbf{I})) + \exp(\text{App}(Y_l; \mathbf{I}))}.$$

We estimate the parameters of the reward model via maximum likelihood by formulating the problem as a binary classification with a negative log-likelihood loss: $L_{\text{MLE}}(\text{App}, \mathcal{D}_{\text{app}}) =$

$-\mathbb{E}_{(Y_w,Y_l;\mathbf{I})\sim\mathcal{D}_{\text{app}}}\log\sigma(\text{App}(Y_w;\mathbf{I})-\text{App}(Y_l;\mathbf{I}))$. To reduce variance, we normalize this by subtracting the population mean so that $\mathbb{E}_{(Y,\mathbf{I})\sim\mathcal{D}_{\text{app}}}[\text{App}(Y;\mathbf{I})]=0$ for all contexts $\mathbf{I}$.

Personalization. A conversational RS is only effective to the extent that it recommends, and ultimately, the user accepts, items of significant value to the user. Thus, *personalization* is perhaps the foremost criterion with which to evaluate an LM-based RS. Particularly, we wish to evaluate the extent to which the LM's generated response Y corresponds to an item with high utility for a user u. To this end, we develop a scoring model $\text{Per}(Y;\mathbf{i},\mathbf{u})$ which interprets the semantics of text Y to quantify its value as a personalized recommendation.

To achieve this, recall the dataset $\mathcal{D}=\{(\mathbf{I}^{(k)},\mathbf{i}^{(k)},\mathbf{u}^{(k)},Y^{(k)})\}_{k=1}^{|\mathcal{D}|}$ of item description, item CF embedding vector, user CF embedding vector, and textual response tailored to the user, and the estimated utility that is the dot product $\hat{r}=\mathbf{i}\cdot\mathbf{u}$ of their CF embedding vectors. To measure personalization one could learn a reward model $\text{Per}(Y;\mathbf{i},\mathbf{u})$ that predicts the utility $\hat{r}$ based on textual response Y. However, this approach relies on a strong assumption that such text alone is predictive of user-item utility. Alternatively, we can also employ the LAIF approach (Christiano et al., 2017) that leverages preference feedback to learn a personalization reward model. Using the same dataset $\mathcal{D}$, and assuming the recommendation text is more personalized than item description, i.e., $Y\succ\mathbf{I}|\mathbf{i},\mathbf{u}$,[2] a Bradley-Terry based personalization reward model $\text{Per}(Y;\mathbf{i},\mathbf{u})$ can be learned by minimizing the negative log-likelihood loss: $L_{\text{MLE}}(\text{Per},\mathcal{D}_{\text{per}})=-\mathbb{E}_{(Y,\mathbf{I};\mathbf{i},\mathbf{u})\sim\mathcal{D}_{\text{per}}}\log\sigma(\text{Per}(Y;\mathbf{i},\mathbf{u})-\text{Per}(\mathbf{I};\mathbf{i},\mathbf{u}))$.

Preference Relevance. While appeal and personalization distinguish compelling recommendations for a user from simple factual item summaries, they do not capture the full *relevance* of the LM's response w.r.t. a user's preferences. For example, the LM might still describe item attributes that the user has no interest in (positively or negatively). To address this, we assume access to a *textual description* of a user's preferences (we later describe how we create these from user CF embeddings). We train an additional reward model, $\text{Prel}(Y;\mathbf{I},\mathbf{u})$, which explicitly measures the semantic similarity between a user's description of preferences and the generated text, constrained to attributes of the recommended item. More specifically, we assume availability of a mapping from a user's CF embedding vector $\mathbf{u}$ to a textual description of their preferences. We train this mapping using a dataset of user embeddings and textual descriptions $\{\text{U}_j(\mathbf{u})\}_{j=1}^{J}$ (see Appendix C for details on the generation of this dataset).

Next, for each $(\mathbf{I},\mathbf{u},Y)$, we encode the user's textual preferences $\{\text{U}_j(\mathbf{u})\}_{j=1}^{J}$ and the item description $\mathbf{I}$ using an *LM semantic encoder*.[3] Then, we rank each textual preference using cosine similarity of its encoded counterpart and encoded item. This, in turn, determines which of the J preference texts are most relevant to the item of interest. Finally, we use the same model to encode the recommendation response Y and compute its cosine similarity with the user preference texts.

We define the *preference relevance* score s of Y w.r.t. user-item pair $(\mathbf{u},\mathbf{i})$ to be the average of the above cosine similarity scores. To this end, we train the reward model $\text{Prel}(Y;\mathbf{I},\mathbf{u})$ by minimizing an ℓ_2 regression loss $L_{\text{REG}}(\text{Prel},\mathcal{D}_{\text{Prel}})=\mathbb{E}_{(\mathbf{I},\mathbf{u},Y,s)\sim\mathcal{D}_{\text{Prel}}}(s-\text{Prel}(Y;\mathbf{I},\mathbf{u}))^2$.

4 Reinforcement Learning based Fine-tuning

RL from AI feedback (RLAIF) can effectively align LMs to metrics that are labeled by off-the-shelf LMs in lieu of humans. Recent work (Lee et al., 2023; Bai et al., 2022; Zhu et al., 2023) has shown that hybrid human-AI preference models, together with *self-improving* fine-tuning, outperforms traditional supervised fine-tuned baselines and offers additional benefits relative to standalone RL fine-tuning with human feedback (RLHF). Using the four principles for LM-based recommendation outlined in Section 3, we develop four reward models to help train and evaluate LM w.r.t. personalization, precision, appeal and preference relevance. We then devise an RLAIF technique to fine-tune an LM with a joint reward model defined by these four components.

In multi-objective RL, it is common to aggregate reward models via *linear scalarization* (Peschl et al., 2021) (which corresponds to solving for an optimum on the convex Pareto frontier). Given a text

[2] Instead of comparing the recommendation text with item description, one could instead construct a dataset with two texts and a labeled rating order (see Appendix C for details).

[3] Much like *Sentence-T5* (Ni et al., 2022a) and *T5-based Retrievers* (Ni et al., 2022b), the semantic encoder E maps textual inputs (e.g., item description $\mathbf{I}$ or user preference texts $\{\text{U}_j(\mathbf{u})\}_{j=1}^{J}$) to a latent space in $\mathbb{R}^{d_{\text{enc}}}$.

response $Y = \{y_n\}_{n=0}^{N-1}$, item description $\mathbf{I}$, and user-item CF embedding vectors $(\mathbf{u}, \mathbf{i})$, we define the LM-based RS reward recommender reward by:

$$r(y_n; y_{0:n-1}; \mathbf{I}, \mathbf{i}, \mathbf{u}) = \begin{cases} \eta_1 \text{NLI}(Y; \mathbf{I}) + \eta_2 \text{App}(Y; \mathbf{I}) + \eta_3 \text{Per}(Y; \mathbf{i}, \mathbf{u}) + \eta_4 \text{Prel}(Y; \mathbf{I}, \mathbf{u}) & \text{if } y_n = [\text{EOS}]; \\ 0 & \text{otherwise,} \end{cases}$$

where $\eta_1, \eta_2, \eta_3, \eta_4 \geq 0$ are importance weights for the component rewards, and are treated as hyper-parameters (optimized using e.g., grid search).

Recall the LM $\mathbb{P}_\theta(Y \mid y_0; \mathbf{I}, \mathbf{i}, \mathbf{u})$ with item text $\mathbf{I}$, item-user CF embedding vectors $(\mathbf{i}, \mathbf{u})$ and the reward model $r(Y, \mathbf{I}, \mathbf{i}, \mathbf{u})$, which jointly measures appeal, factuality, preference-relevance, and personalization of a recommendation response. The goal in LM fine-tuning is to maximize the average overall quality of the generated text, i.e., $\max_\theta \ \mathbb{E}_{(\mathbf{I},\mathbf{i},\mathbf{u})} \mathbb{E}_{\mathbb{P}_\theta(Y|\mathbf{I},\mathbf{i},\mathbf{u})}[r(Y; \mathbf{I}, \mathbf{i}, \mathbf{u})]$. Using the CoMDP framework, it is easily shown that this learning problem can be solved with on-policy REINFORCE (Williams, 1992), in which the policy gradient is estimated using trajectories generated by the current LM policy.

A risk of RL fine-tuning based on an AI-feedback is that it might overfit to the model, thereby degrading the "skill" of the original LM. To alleviate this, we add a KL regularization term (Ouyang et al., 2022; Stiennon et al., 2020) between the LM $\mathbb{P}_\theta(Y|\mathbf{I}, \mathbf{i}, \mathbf{u})$ and the pre-trained model $\mathbb{P}_{\text{pre}}(Y|\mathbf{I}, \mathbf{i}, \mathbf{u})$ to the CoMDP objective function. Leveraging the auto-regressive nature of LMs, KL regularization is applied over the entire MDP trajectory, reducing the objective function to

$$\max_\theta \ J(\theta) := \mathbb{E}_{(\mathbf{I},\mathbf{i},\mathbf{u})} \mathbb{E}_{\mathbb{P}_\theta(Y|\mathbf{I},\mathbf{i},\mathbf{u})} \left[r(Y; \mathbf{I}, \mathbf{i}, \mathbf{u}) - \beta \log \frac{\mathbb{P}_\theta(Y|\mathbf{I}, \mathbf{i}, \mathbf{u})}{\mathbb{P}_{\text{pre}}(Y|\mathbf{I}, \mathbf{i}, \mathbf{u})} \right]. \tag{1}$$

This is equivalent to a KL-regularized CoMDP. The LM policy π_θ, where $\mathbb{P}_\theta(Y|\mathbf{I}, \mathbf{i}, \mathbf{u}) = \prod_{n=0}^{N-1} \pi_\theta(s_n|a_n; c)$, can be learned by computing the policy gradient of the KL-regularized objective online, or by employing an off-policy RL algorithm, e.g., SAC (Haarnoja et al., 2018), in-sample softmax (Xiao et al., 2023), CQL (Kumar et al., 2020), that leverages offline data $\mathcal{D}$ for more efficient training. (See Appendix D for full exposition of these algorithms.) KL regularization, intended to avoid over-fitting to the reward model, can also alleviate out-of-distribution generalization issues common in offline RL (Kumar et al., 2019).

5 Experiments

We conduct empirical validations of P^{4}LM, focusing on assessing its capability to generate factual, personalized, and compelling recommendation endorsements. We examine the hypothesis that the reward models (RM) detailed in Section 3 significantly increase the personalization, precision, appeal and preference relevance of movie recommendations. We use the MovieLens 25M recommendation dataset (Harper & Konstan, 2015), which contains ratings of $62,423$ movies by $162,541$ users. We use these movie-user interactions to generate movie descriptions, user-preference texts, and sample recommendation responses by prompting a PaLM2-L LM (Google et al., 2023); our data generation procedures are detailed in Appendix C. The resulting datasets have four components: (1) movie descriptions $\mathbf{I}$, (2) item-user behavioral embeddings $(\mathbf{i}, \mathbf{u})$, (3) user preference texts $\text{U}(\mathbf{u})$, and (4) sample responses Y. We experiment with a set of LMs in the PaLM2 family (Google et al., 2023). To incorporate user and movie embedding vectors into the LM (Section 3) we construct LMs by augmenting these LMs with adapter layers. Specifically, we train two models, P^{4}LM and P^{4}LM-S, derived from PaLM2-XS and PaLM2-XXS, respectively. Our reward mixing weights, optimized using grid search, are $(\eta_1, \eta_2, \eta_3, \eta_4) = (2.0, 0.1, 1.0, 1.0)$.

To demonstrate the efficacy of our models P^{4}LM and P^{4}LM-S, we compare them with the following SOTA baselines on our conversational movie recommendation task: (i) **PaLM2-L**, a pre-trained model prompted using movie descriptions, user preference texts and instructions to generate a response that respects our four recommender principles; (ii) **Supervised Fine-Tuned with Text (SFT-Text)**, a PaLM2-XS model fine-tuned with the dataset above, with explicit user-item texts as input; (iii) **Supervised Fine-Tuned (SFT)**, a PaLM2-XS model fine-tuned to use user-item embedding vectors.

In Section 5.1, we first validate the efficacy of the RMs using AI-generated examples with known labels. In the following sections, we measure the performance of our approach via *model-based* and *human* evaluation.

Table 2: Model-based Evaluation Based on the Principles of Recommendation LM.

Method	Precision	Personalization	Appeal	Pref. Relevance
PaLM2-L	0.52 ± 0.03	-0.04 ± 0.04	0.36 ± 0.04	-
SFT-Text	0.58 ± 0.03	0.10 ± 0.04	0.46 ± 0.03	$\mathbf{-1.01 \pm 0.06}$
SFT	0.58 ± 0.03	-0.15 ± 0.04	0.33 ± 0.04	-1.08 ± 0.07
P^{4}LM	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.23 \pm 0.04}$	0.63 ± 0.04	-1.18 ± 0.07
P^{4}LM-S	0.65 ± 0.03	0.18 ± 0.04	$\mathbf{0.72 \pm 0.04}$	-1.10 ± 0.08

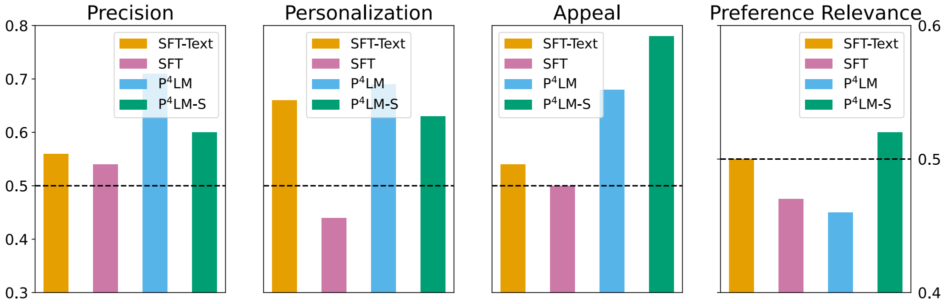

Figure 2: Win Rates of Different Model-based Scores against PaLM2-L

5.1 Validation of Reward Models

Table 1: The accuracies of the RMs on GPT-4 generated examples.

RM	NLI	App	Per
Accuracy	1.0	1.0	0.9375

It is crucial to assess if the scores of the RMs reflect human values. To test this, we prompt an LM with $(\mathbf{I}, U(\mathbf{u}))$, generating two distinct-quality endorsements. For instance for NLI, the LM produces one response adhering to the item description and another with added hallucinations. The first response is more factually accurate. With n examples, we evaluate the NLI RM's accuracy in distinguishing better from worse endorsements in precision. This process is repeated for App and Per RMs, with accuracies presented in Table 1. For robustness in evaluation, we use GPT-4 (OpenAI, 2023) for generating the test data. See Appendix A.1 for additional details.

5.2 Model-Based and Human Evaluation

Model-Based Evaluation We conduct *model-based* evaluations using the four criteria and their corresponding RMs from Section 3, namely, NLI, App, Per, and Prel. Specifically, we report the scores of responses Y from different models on a held-out, unlabeled dataset $\mathcal{D}_{\text{test}}$ consisting of 96 non-overlapping user-movie pairs. We also assess the relative improvement of each LM over the PaLM2-L common baseline that we use to generate the supervised learning dataset. We do this by computing the (a) *win rate* (number of occurrences on which a candidate LM outperforms PaLM2-L), (b) *absolute increase* (the magnitude of the score improvement), and (c) *percentage increase* (the relative score improvement).[4]

Our results in Table 2 highlight the robust performance of P^{4}LM in three pivotal dimensions: Precision, Personalization, and Appeal. P^{4}LM attains the highest precision score by a wide margin, underscoring its ability to mitigating the risks of misleading users with hallucinated information about recommended items. It also outperforms on personalization and appeal. It appears that P^{4}LM compromises on preference relevance to achieve these gains, with qualitative comparisons (see Appendix A.2 for details) on the texts generated by P^{4}LM and SFT verifying these phenomenons. We believe that personalization is by far the most important aspect of recommendation quality, while precision/factuality is the most critical property of any endorsement text.

Figure 2 displays the win rates of various LMs compared to PaLM2-L.[5] Notably, SFT and SFT-Text show low precision scores, indicating a tendency to overfit and generate inaccurate movie details.

[4] Precise definitions of these relative metrics are provided in Appendix B.

[5] See Appendix A for detailed absolute and percentage increases in Figures 4 and 3.

Table 3: Human Evaluation Results: Ratio of Preferred Outputs over PaLM2-L Outputs

Method	Precision	Personalization	Appeal	Pref. Relevance
SFT-Text	0.29	0.458	0.375	0.458
SFT	0.458	**0.583**	0.458	0.583
P^{4}LM	**0.5**	**0.583**	**0.667**	**0.708**
P^{4}LM-S	**0.5**	0.5	0.5	0.333

Table 4: Model-based Evaluation Scores using a Single Reward Model (Ablation Studies)

Method	Precision	Personalization	Appeal	Pref. Relevance
NLI	$\mathbf{0.77 \pm 0.02}$	-0.14 ± 0.04	0.43 ± 0.04	-1.08 ± 0.07
Personalization	0.55 ± 0.03	$\mathbf{0.39 \pm 0.04}$	$\mathbf{0.94 \pm 0.04}$	-1.11 ± 0.07
Appeal	0.56 ± 0.03	0.14 ± 0.04	0.76 ± 0.04	-1.06 ± 0.06
Pref. Relevance	0.51 ± 0.03	-0.07 ± 0.04	0.51 ± 0.04	$\mathbf{-1.02 \pm 0.07}$

P^{4}LM's higher personalization score, even surpassing SFT-Text that directly uses user profile texts, highlights its superior utilization of user behavioral information in the user embedding vector. Unlike SFT, P^{4}LM effectively leverages this data, emphasizing the advantage of RL-training.

Human Evaluation Table 3 shows our human evaluation results. We presented *raters* with two endorsement texts: one from PaLM2-L and another by a different model, to assess their relative performance. Raters evaluated these texts based on four criteria aligned with our model-based evaluation. For more information, refer to Appendix E.

In precision, P^{4}LM and P^{4}LM-S equaled PaLM2-L in raters' preferences, despite their smaller sizes (XS and XXS). P^{4}LM also outperformed baseline models in personalization, appeal, and preference relevance. Interestingly, human evaluation favored SFT over SFT-Text, contrary to the model-based results. This may be because SFT-Text, sharing the same inputs as PaLM2-L and trained on its generated dataset, is limited by PaLM2-L's capabilities. Conversely, SFT's use of user embeddings enhances personalization, giving it an edge over SFT-Text. Overall, the human evaluation further corroborates the results of the model-based assessment, highlighting P^{4}LM's superior performance.

Ablation Studies Our ablation studies, outlined in Table 4, show that training with a single RM predictably results in policies scoring the highest in the model-based evaluation for the specific RM targeted (see Precision, Personalization, Preference Relevance scores). Intriguingly, a model trained solely on Personalization not only excels on that metric, but also attained the highest score in Appeal, suggesting a possible correlation where recommendation text that is well-tailored to a user's preferences may be inherently appealing. We also explore the impact of varying the mixing-weight combination, which is discussed in Appendix A.

We have conducted human evaluations on models trained with individual RMs, with results being detailed in Table 5 in Appendix A. Here, we briefly highlight key observations, directing readers to the appendix for in-depth discussion. Notably, the performance ranking of these models in human evaluations diverges from those in Table 4. The model trained solely with the NLI reward, expected to excel in Precision, surprisingly scored the lowest in human evaluations. This suggests potential *reward hacking*, where the policy learner exploits the single RM to inflate scores. This underscores the necessity of using multiple RMs, as each serves as a regularizer thwarting over-optimization of a single RM, ensuring balanced performance. Our ablation studies reveal that focusing on a particular RM boosts its model-based score, but this trend is not always reflected in human evaluations, further indicating the possibility of reward hacking. This stresses the importance of adopting a diverse set of RMs in RLAIF to counteract such issues.

6 Related Work

Our work intersects multiple areas of research, notably personalized recommendation systems, leveraging of language models (LMs) and reinforcement learning, recommendation integrity.

Personalized Recommender Systems Recommender systems have ubiquitous applications permeating e-commerce, content providers, social media, etc., with collaborative filtering (CF) (Schafer et al., 2007) as the prominent modeling technique. Early works include matrix factorization approaches (Mnih & Salakhutdinov, 2007), which became a foundation for subsequent deep learning methods like neural CF (He et al., 2017). Notably, dual encoder architectures emerged, where user and item embeddings are co-trained (Yi et al., 2019; Yang et al., 2020). While traditional CF approaches worked well in many applications, advances in deep personalization allow user and item embeddings to capture more nuanced preferences (Rendle et al., 2020; Beutel et al., 2018).

Conversational Recommender Systems & Language Models Conversational recommender systems (RSs) add an interactive layer over traditional RSs with an conversational agent interacting with users, understanding their preferences and refining recommendations through dialogue (Chen et al., 2019; Zhou et al., 2020; Lei et al., 2020; Li et al., 2018; Sun & Zhang, 2018; Christakopoulou et al., 2016). This paradigm integrates aspects of natural language understanding, making it ripe for integrating LMs. Leveraging language models in RSs is a relatively recent development. With the advance of transformer architectures (Vaswani et al., 2017; Wolf et al., 2019), LMs have found use-cases beyond typical NLP tasks. Researchers began exploring the synthesis of textual data with user preferences to enhance the personalization and expressiveness of RSs (Jaech & Ostendorf, 2018; Xia et al., 2023). Our work situates itself in this space, but with an added twist: we aim to generate compelling narratives that genuinely communicate the relevance of a recommendation.

Transparency and Truthfulness in Recommendation Systems Maintaining integrity in RSs is technically challenging yet critically important. The potential that RS algorithms inadvertently mislead users or reinforce biases has been highlighted (Abdollahpouri et al., 2019; Shen et al., 2023; Cabello et al., 2023). Therefore, increasingly researchers are not only prioritizing the recommendation efficacy but also the fairness, transparency, and interpretability of RS algorithms (Beutel et al., 2019; Ghazimatin et al., 2020; Chen et al., 2023). Our work takes cues from this domain, emphasizing truthful and precise recommendations that articulate genuine merits rather than compelling distortions.

Reinforcement Learning with Human/AI Feedback The integration of reinforcement learning (RL) with language models has emerged as a compelling strategy for refining model behavior beyond supervised fine-tuning (Williams, 1992; Ranzato et al., 2016). The RL with Human Feedback (RLHF) methodology (Christiano et al., 2017; Bai et al., 2022), in particular, has gained traction, where model responses are ranked by human evaluators and subsequently used to fine-tune models through techniques like Proximal Policy Optimization (Schulman et al., 2017). In a different vein, Inverse Reinforcement Learning (Abbeel & Ng, 2004) has been employed to extract objectives from expert demonstrations in textual settings (Daniels-Koch & Freedman, 2022; Sun, 2023). Additionally, there's a growing interest in AI-driven feedback mechanisms, where preferences are labeled by off-the-shelf LMs in lieu of humans (Lee et al., 2023; Bai et al., 2022). These endeavors underline the potential of using RL to steer LMs towards better alignment with human preferences and nuanced task objectives.

7 Conclusion

We studied language modeling for personalized recommendation. By developing novel reward models which quantify prominent attributes of personalized recommendations, one may develop self-improving LM methodologies via reinforcement learning with AI feedback. As a result, our developed LM; namely P^{4}LM, not only parses language semantics, but also understands latent user preferences (encoded in the CF embedding space). P^{4}LM provides factual, compelling, personalized endorsement of relevant items, connecting the items with users' preferences, thereby increasing the likelihood of users accepting high-value recommendations.

We demonstrated the efficacy of P^{4}LM on the MovieLens 25M dataset. Particularly, our agent better understands user behaviors encoded in the CF embedding space and delivers precise, compelling, personalized movie recommendation narratives. Our work is a step toward creating intelligent conversational recommenders which can compellingly explain the intricacies between item features and user preferences. Future work includes (i) improving P^{4}LM's capabilities to generate longer responses beyond standard single-shot autoregressive decoding; (ii) extending our RL fine-tuning approach to handle multi-turn conversational recommendations; (iii) developing better reasoning capabilities to trade off between user-item preferences and constraints; (iv) and expanding the LM's functionality beyond recommendation, to also include technical support, negotiations, etc.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The impact of popularity bias on fairness and calibration in recommendation. *arXiv preprint arXiv:1910.05755*, 2019.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

Kavosh Asadi and Jason D Williams. Sample-efficient deep reinforcement learning for dialog control. *arXiv preprint arXiv:1612.06000*, 2016.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 46–54, 2018.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2212–2220, 2019.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 370–378, 2023.

Salvatore Carta, Anselmo Ferreira, Alessandro Sebastian Podda, Diego Reforgiato Recupero, and Antonio Sanna. Multi-dqn: An ensemble of deep q-learning agents for stock market forecasting. *Expert systems with applications*, 164:113820, 2021.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391*, 2019.

Jen-Tzung Chien and Che-Yu Kuo. Markov recurrent neural network language model. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 807–813. IEEE, 2019.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 815–824, 2016.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1331–1340. PMLR, 2019.

Oliver Daniels-Koch and Rachel Freedman. The expertise problem: Learning from specialized feedback. *arXiv preprint arXiv:2211.06519*, 2022.

Alexander Felfernig, Gerhard Friedrich, Bartosz Gula, Martin Hitz, Thomas Kruggel, Gerhard Leitner, Rudolf Melcher, Daniela Riepan, Sabine Strauss, Erich Teppan, et al. Persuasive recommendation: serial position effects in knowledge-based recommender systems. In *Persuasive Technology: Second International Conference on Persuasive Technology, PERSUASIVE 2007, Palo Alto, CA, USA, April 26-27, 2007, Revised Selected Papers 2*, pp. 283–294. Springer, 2007.

Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*, 2023.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. Prince: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 196–204, 2020.

Rohan Anil Google, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 technical report, 2023.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

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

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pp. 173–182. ACM, 2017.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. Q{2}: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*, 2021.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. pp. 3905–3920, July 2022.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Tzu-Kuo Huang, Ruby C Weng, and Chih-Jen Lin. Generalized bradley-terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7(1), 2006.

Aaron Jaech and Mari Ostendorf. Personalized language model for query auto-completion. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 700–705. Association for Computational Linguistics, 2018.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. *arXiv preprint arXiv:2108.13684*, 2021.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. Conversational recommendation: Formulation, methods, and evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2425–2428, 2020.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31, 2018.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*, 2022.

Bill MacCartney and Christopher D Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 193–200, 2007.

Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20, 2007.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained Text-to-Text models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 1864–1874. Association for Computational Linguistics, 2022a.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 9844–9855, 2022b.

OpenAI. Gpt-4 technical report, 2023.

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

Markus Peschl, Arkady Zgonnikov, Frans A Oliehoek, and Luciano C Siebert. Moral: Aligning ai with human norms through multi-objective reinforced active learning. *arXiv preprint arXiv:2201.00012*, 2021.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, 2016.

Steffen Rendle, Walid Krichene, Li Zhang, and John R. Anderson. Neural collaborative filtering vs. matrix factorization revisited. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, pp. 240–248. ACM, 2020.

Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*, 2023.

Fumihiro Sasaki and Ryota Yamashina. Behavioral cloning from noisy demonstrations. In *International Conference on Learning Representations*, 2020.

J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pp. 291–324. Springer, 2007.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Tianshu Shen, Jiaru Li, Mohamed Reda Bouadjenek, Zheda Mai, and Scott Sanner. Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. *Information Processing & Management*, 60(1):103139, 2023.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

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

Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

Hao Sun. Offline prompt evaluation and optimization with inverse reinforcement learning. *arXiv preprint arXiv:2309.06553*, 2023.

Yueming Sun and Yi Zhang. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*, pp. 235–244, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xue Xia, Pong Eksombatchai, Nikil Pancha, Dhruvil Deven Badani, Po-Wei Wang, Neng Gu, Saurabh Vishwas Joshi, Nazanin Farahpour, Zhiyuan Zhang, and Andrew Zhai. Transact: Transformer-based realtime user action model for recommendation at pinterest. *arXiv preprint arXiv:2306.00248*, 2023.

Chenjun Xiao, Han Wang, Yangchen Pan, Adam White, and Martha White. The in-sample softmax for offline reinforcement learning. *arXiv preprint arXiv:2302.14372*, 2023.

Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H Chi. Mixed negative sampling for learning two-tower neural networks in recommendations. In *Proceedings of the Web Conference (WWW-20)*, pp. 441–447, Taipei, 2020.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the Thirteenth ACM Conference on Recommender Systems (RecSys19)*, pp. 269–277, Copenhagen, 2019.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pp. 11328–11339. PMLR, 2020a.

Yongfeng Zhang, Xu Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020b.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*, 2020.

Banghua Zhu, Jiantao Jiao, and Michael I Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.