【LLM论文阅读】RLRF4Rec：从推荐系统反馈中进行强化学习以增强推荐重排序

原创 方方 方方的算法花园 2024年11月04日 11:04 北京

0 论文概况

1. 论文名称：
RLRF4Rec: Reinforcement Learning from RecsysFeedback for Enhanced Recommendation Reranking 《RLRF4Rec：从推荐系统反馈中进行强化学习以增强推荐重排序》
2. 论文链接：https://arxiv.org/pdf/2410.05939
3. 论文作者所在机构：北京大学智能科学与技术学院、通用人工智能国家重点实验室
4. 一句话概括：这篇论文提出了 RLRF4Rec 框架，通过让大语言模型生成用户偏好知识来增强传统推荐模型，利用推荐反馈训练大语言模型，采用 DPO 优化，经实验验证其在推荐重排序任务中显著提升性能。

1 研究背景

1.大语言模型的能力优势：大语言模型具备先进的自然语言理解能力，能够高精度地理解和生成类人文本，展现出令人瞩目的推理能力，可执行复杂认知任务并提供有洞察力的回复。
2.在推荐系统中的应用现状与分类：现有方法主要分为两类：一是利用 LLMs 直接进行推荐，将推荐任务格式化为自然语言结构以微调 LLMs；二是使用 LLMs 增强传统推荐模型，通过对齐潜在表示将语义知识从 LLMs 转移到协作模型，以提升推荐性能。
3. 面临的问题与挑战：尽管 LLMs 知识丰富，但由于其**预训练任务与推荐任务差异较大，且预训练时推荐数据不足，导致其在推荐任务中的性能有待提高。**例如，当前基于 LLMs 的知识增强推荐多采用零样本方式，存在诸如如何使 LLMs 与推荐系统更好地对齐，以及如何用自然语言表示协作关系并融入推荐模型训练等问题。

2 论文贡献点

为了应对这些挑战，作者提出了 RLRF4Rec，一个用于调整大型语言模型以适应推荐系统的框架。首先让LLM根据用户交互历史生成推断的用户偏好，然后将其用于增强传统的基于ID的序列推荐模型。随后，基于 Kar 设计了一个奖励模型，以评估大型语言模型推理知识的质量。然后，从 N 个样本中选择最佳和最差响应来构建LLM调优的数据集。最后，使用DPO设计了一种结构对齐策略，以便仍然能够从LLM获得良好的推理知识。总的来说，贡献总结如下：
（1）定义推荐系统新问题：将LLM与知识增强推荐进行对齐，揭示了基于上下文学习方法的局限性，并强调了指令调优的重要性。
（2）提出了一种双向方法，允许推荐模型向LLM提供反馈，使其与推荐任务对齐。
（3）选择了两个数据集，并使用两个模型进行了实验，验证了所提出框架的有效性和高效性。该方法与基线相比取得了性能提升。
(KAR：https://github.com/YunjiaXi/Open-World-Knowledge-AugmentedRecommendation/tree/main/data/amz/knowledge)

3 RLRF4Rec 具体方案

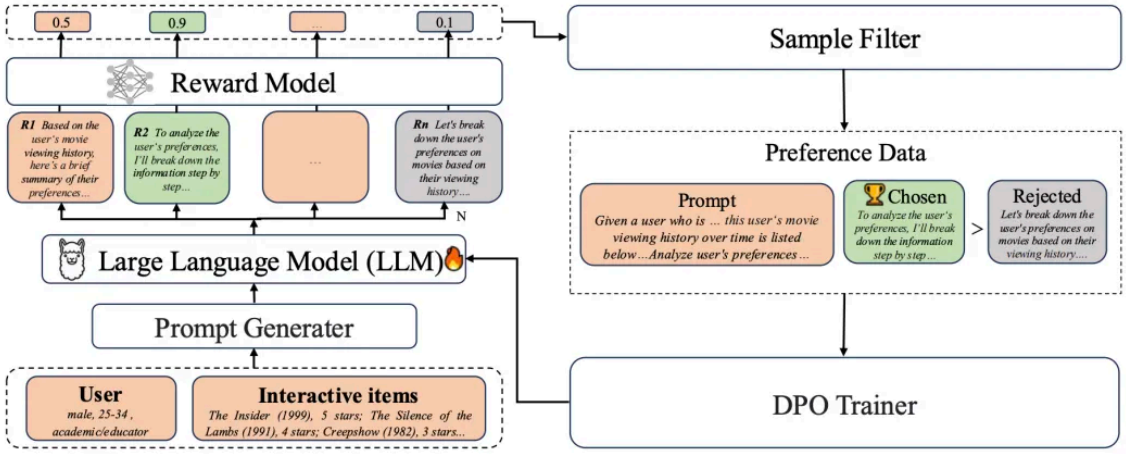

Fig. 2. Illustration of the RLRF4Rec framework constructed by three stages: knowledge augmentation, recommendation feedback, and knowledge alignment .For the knowledge augmentation stage, LLM generateste N responses from recommendation data. For the recommendation feedback stage, an existing recommendation model is trained as a reward model. For the knowledge alignment stage, the best and worst responses are selected by the reward model, then we implement DPO methods to get the final LLM.

RLRF4Rec 是一种用于增强推荐系统中重排序任务的框架，具体方案如下：

1.知识增强（Knowledge Augmentation from LLM）

（1）任务分解与推理知识生成：首先将复杂的推理任务分解，确定影响用户偏好和物品特征的主要因素。然后针对每个因素，利用公开预训练模型（如 LLaMA 3.1）生成关于用户偏好的推理知识。不使用 LLM 生成物品的事实性知识，而是利用来自 KAR 的公共资源。下图为如何生成用户偏好的prompt：

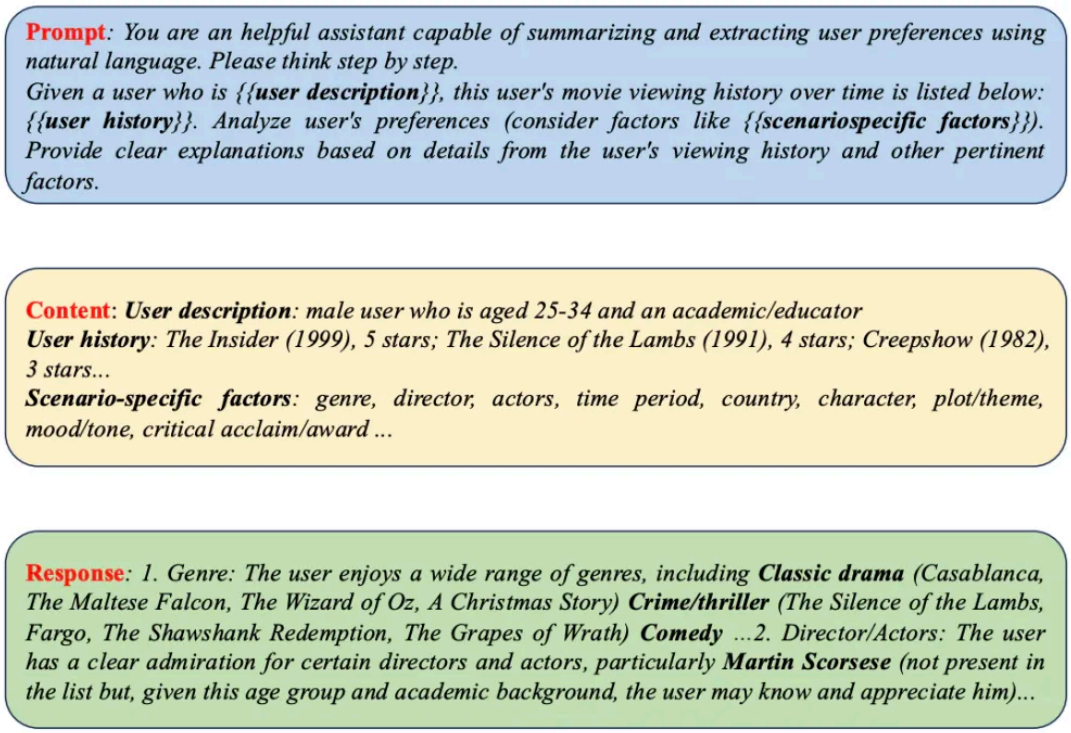

Fig. 3. A Prompt example aimed at extracting descriptive knowledge and user preferences about items using LLMs. The prompt template is represented by the blue bubble. A prompt can be generated by injecting specific information about the user content to be filled in the template shown in the yellow bubble into the prompt template. This prompt then instructs the LLM to generate descriptive knowledge and deduce user preferences, as indicated in the green bubble.

（2）向量转换：通过一个适配器将生成的推理和事实性知识转换为低维向量，使其与传统推荐模型兼容，以便后续利用这些知识增强推荐模型。

2. 推荐反馈（Feedback of the Recommendations）

（1）评估推理知识质量：通过推荐性能来评估 LLM 生成的推理知识的质量。具体使用如 NDCG/MAP 等可针对单个用户测量的指标，从 LLM 生成的 N 个样本中为每个用户选择最佳和最差的推理知识样本。

（2）构建微调数据集：将最佳和最差样本分别作为推荐偏好中的选择和拒绝样本，构建微调数据集。

（3）强化学习训练 LLM：应用直接偏好优化（DPO）等强化学习方法训练 LLM。DPO 的关键创新是避免显式奖励建模，直接根据推荐偏好优化语言模型。具体通过定义偏好目标函数，将人类偏好直接转换为训练使用的损失函数，公式为

$$L_{DPO} = -log\sigma\left(\beta log\frac{\pi\left(y_1|x\right)}{\pi_{ref}\left(y_1|x\right)} - \beta log\frac{\pi\left(y_2|x\right)}{\pi_{ref}\left(y_2|x\right)}\right)$$

，其中πref是基础参考模型，σ 是 sigmoid 函数，β 是超参数（默认值为 0.01）。

3. 知识对齐（Knowledge Alignment）

（1）生成响应与评估：采用多生成方法，让 LLM 生成 N 个响应，然后使用预训练的推荐模型作为奖励模型来评估这些响应的质量。

（2）选择样本与训练：从 N 个响应中选择最佳和最差响应作为选择和拒绝示例，再使用 DPO 方法训练 LLM，得到最终基于 LLM 的推荐模型。其详细学习过程如 Algorithm 1 所

示，通过迭代优化，使 LLM 与推荐模型的能力更好地对齐，从而提高推荐系统的性能。

Algorithm 1 Iterative optimization method

Require: Number of responses N, pre-trained recommendation model Rec, LLM with parameters θ_{LLM}
Ensure: Final LLM-based recommendation model
Generate N responses using the LLM:
for $i = 1$ to N **do**
$response_i \leftarrow LLM.generate(prompt_i)$
end for
Evaluate each response using the pre-trained recommendation model:
for $i = 1$ to N **do**
$score_i \leftarrow Rec.evaluate(response_i)$
end for
Select the best and worst responses:
$best\_response \leftarrow response_{\arg\max(score)}$
$worst\_response \leftarrow response_{\arg\min(score)}$
Train the LLM using DPO with the best and worst responses:
$LLM.train(prompt_i, best\_response, worst\_response)$
return Final LLM-based recommendation model

4 实验设置及结论

1.实验设置

（1）数据集：

MovieLens - 1M：包含 6040 个用户对 3416 部电影的 100 万条评分数据。将评分转换为二元标签（4 分和 5 分为正，其他为负），按用户 ID 将数据划分为训练集（90%）和测试集（10%）。

Amazon - Books：是亚马逊评论数据集中的“书籍”类别。经过筛选后，剩余 11906 个用户、17332 个物品以及 1406582 次交互。预处理方式与 MovieLens - 1M 类似，但无用户特征。

（2）baseline模型：在重排序任务中实现了两个最先进的模型，即 DLCM 和 PRM 作为baseline模型。DLCM 首先使用 GRU 对顶部结果进行编码和重排序；PRM 则采用自注意力机制来建模任意物品对之间的相互影响以及用户偏好。

（3）评估指标：在重排序任务中，采用广泛使用的 NDCG@K 和 MAP@K 作为评估指标，用于评估从用户未交互的整个物品集中选择的前 K 个物品。

（4）实现细节：

- **LLM 选择与微调：**使用 Llama - 3.1 - 8BInstruct 作为 RLRF4Rec 的 LLM，并使用 LLaMA - Factory 框架对其进行微调，该框架可适用于其他 LLM。
- **推荐模型训练：**在训练推荐模型作为奖励模型时使用一致参数，以确保与基线模型进行公平比较。
- **奖励模型训练**：利用 Hugging Face 发布的 BERT 对知识进行编码，实验参数设置为 batch size = 64、learning rate = 1e - 3、temperature = 0.03、epoch = 50。
- **DPO 训练：**参数设置为 batch size = 2、learning rate = 5e - 5、gradient accumulation steps = 8、epoch = 3.0。

2.结论

（1）LRF4Rec 有效性验证：通过大量实验验证了 RLRF4Rec 框架的有效性，在推荐重排序指标上取得显著改进。在 MovieLens - 1M 和 Amazon - Books 数据集上，对于不同baseline模型（DLCM 和 PRM），RLRF4Rec 在 NDCG@5 和 MAP@5 等指标上的得分均高于基线模型（base 和 KAR 模型）。

（2）组件重要性分析

知识组件关键作用：消融实验表明知识组件对推荐有效性至关重要，去除知识组件会导致性能显著下降，尤其是在 Amazon - Books 数据集上。

DPO 组件影响：去除 DPO 组件也会对性能产生一定影响，说明完整模型中各组件的协同作用显著增强推荐性能。

（3）迭代性能表现

在 MovieLens - 1M 数据集上验证了模型在多次迭代后的性能稳定性。经过 2 轮迭代后，NDCG@5 和 MAP@5 指标的性能提升变得微不足道，表明模型效果达到饱和点，超过该阈值后可能无需额外迭代。

综上所述，RLRF4Rec 通过整合强化学习和 LLM，有效提升了推荐系统性能，各组件协同工作，且在适当迭代次数后达到性能瓶颈。

END

强化学习 1　LLM论文阅读 13　LLM与推荐 15