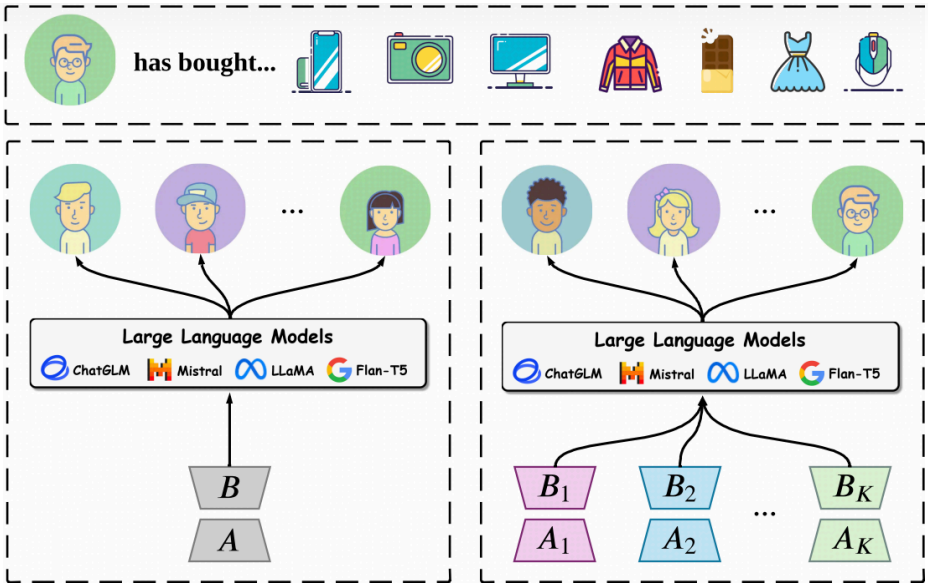
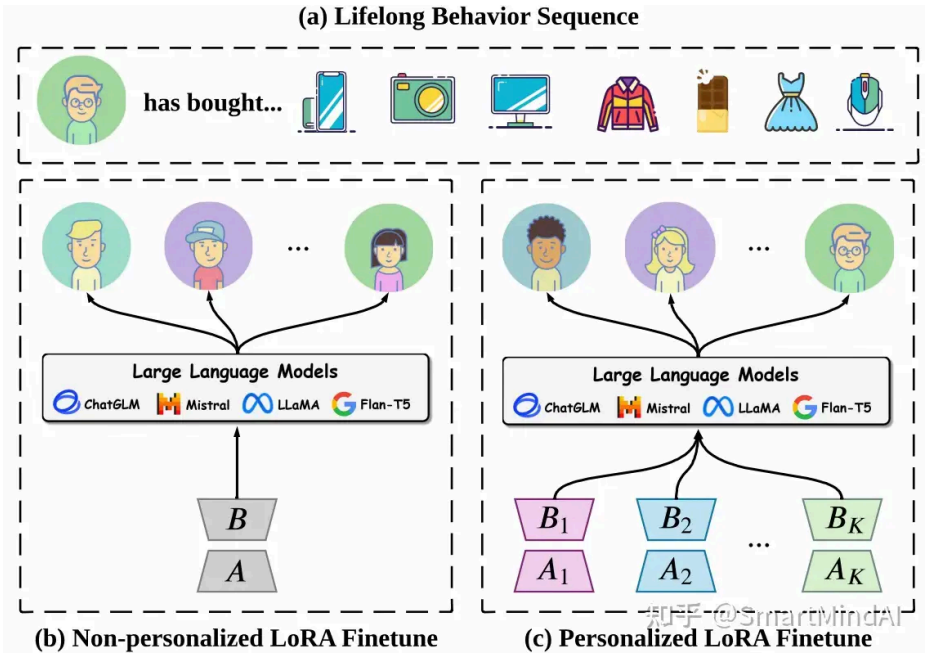


2024华为：RecLoRA-基于LoRA框架的LLM赋能推荐系统的长序列建模

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

Introduction

大型语言模型（LLMs）在自然语言处理（NLP）领域的显著的增长，展现出惊人的能力来理解人类语言，并为各种任务生成与人类写作高度相似的文本。这些模型在处理广泛范围的任务时，通常能够生成与人类写作相似度极高的文本。 近期的研究已经开始探索在不同推荐任务下将 LLMs 应用于推荐系统的可能性。这些研究通常会向 LLMs 注入来自推荐领域的个性化知识。这种个性化知识在适应个人偏好和提升用户互动与参与度方面发挥着关键作用。因此，我们的关注点是提供有用的个性化知识以及如何将其注入到 LLMs 中以提升推荐效果。



主流研究则设计记忆机制来构建全生命周期行为序列模型。其他方法则通过从整个全生命周期序列中检索相关信息来提取关键信息。这些方法验证了用户全生命周期行为序列的有效性，并展示了在基于 LLMs 的推荐模型中的潜力。 当前有研究探讨如何将个性化知识融入大型语言模型（LLMs）。一些工作通过在输入层定制输入模板，将来自传统推荐模型（CRMs）的个性化信息注入到 LLMs 中。其他研究则在预测层通过对语义空间和推荐空间的匹配来实现。

个性化并未得到充分确立。为了提高内存效率，主流方法通常采用低秩适应（LoRA）技术，并进行参数高效的微调（PEFT），将推荐领域的个性化知识注入到大型语言模型（LLMs）中。在PEFT的背景下，为每位用户维护一个独特的LoRA矩阵可以实现最大程度的个性化，有效地捕捉和适应每位用户的独特特征以及兴趣的变化。然而，当前的方法仅实现了提示个性化或表示个性化，而LoRA模块，微调大型语言模型（LLMs）的核心组件，对所有用户和目标项目保持静态状态。这种参数个性化不足导致现代推荐系统在有效管理用户动态和偏好变化方面的能力较差。

其次，在提示或表示层中使用终身个性化行为序列存在有效性与效率的问题。由于大型语言模型（LLMs）对输入长度敏感且有输入大小的限制，它们在处理长行为序列时表现不佳，导致有效性上限。此外，随着输入行为序列长度的增加，训练和推理所需的时间迅速增加，从而导致效率问题。

第三，现有方法在大规模工业数据集上缺乏可扩展性，主要是由于训练效率低下。即使使用LoRA技术，对推荐数据进行微调也会消耗大量的计算资源和时间，对于大规模数据集（通常涉及数百万甚至数十亿条记录）来说，效率极低。虽然有些研究建议在训练数据的小部分（例如少于10%）上微调大型语言模型（LLMs）以平衡推荐性能和训练效率，这种方法限制了大型语言模型（LLMs）对完整训练空间的了解，导致性能不佳。因此，如何在确保训练效率的同时使大型语言模型（LLMs）充分感知训练空间仍然是一个开放的研究问题。

为了解决上述问题，我们提出了一种新的大型语言模型个性化低秩适应框架（推荐系统，简称RecLoRA）。具体来说，我们维护一组并行、独立的LoRA权重，而不是像以往工作建议的那样使用单一静态LoRA模块。如图所示，这个个性化的LoRA模块用于将个性化用户知识融入大型语言模型中，作为参数个性化策略。我们为用户分配一个连续的CRM，并根据CRM适应用户表示到个性化的LoRA权重。通过在连续CRM中为用户分配更长的序列，RecLoRA显著提升了推荐的有效性，而成本的增加几乎可以忽略不计。此外，CRM在完整训练空间上预训练，因此基于CRM动态生成LoRA权重可以像放大镜一样，将从少量数据中观察到的训练信号放大到整个数据空间，而无需增加大型语言模型（LLMs）的训练时间。

Preliminary

Problem Formulation

推荐系统的核心功能是在特定情境下，预测用户对目标项目的偏好。我们将历史数据集表示为：

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

其中N代表数据实例的数量。输入数据x_i包含用户属性u_i（例如，年龄、位置和历史操作），项目特征v_i（例如，类别和品牌），上下文c_i（例如，时间和季节）以及用户历史行为序列$\mathcal{H}_{u_i}$（如点击、评分等操作）：

$$x_i = (u_i, v_i, c_i, \mathcal{H}_{u_i}),$$

其中$\mathcal{H}_{u_i}$表示用户u_i的连续发生的行为序列，包含N_{u_i}个行为。

标签$y_i \in \{1, 0\}$指示了用户对项目进行的偏好交互信号，其中1代表偏好0代表不偏好：

$$y_i = \begin{cases} 1, & u_i \text{ clicks or likes } v_i; \\ 0, & \text{otherwise.} \end{cases}$$

根据不同的输入数据转换，通常存在两种不同的推荐模式：（1）通过独热编码方式获取的CRM输入数据；（2）由硬提示模板生成的，作为推荐系统的LLM（语言模型）输入数据。

Conventional Recommendation Models

对于ID输入x_i^{ID}，不同的传统推荐模型（CRM）被设计用于从不同方面捕捉特征交互和用户行为建模的规律，以准确估计用户偏好。CRM的一般形式如下：

$$h_i^{ID} = \text{CRM}(x_i^{ID}),$$
$$\hat{y}_i^{ID} = \sigma(\text{MLP}(h_i^{ID})) \in (0, 1),$$

知乎

Large Language Models as Recommenders

通常，大型语言模型（LLMs）指的是参数量达到数十亿级别的基于Transformer的语言模型，这些模型在庞大的文本数据集上进行训练，展现了在各种自然语言处理任务中非凡的能力。 在直接采用LLMs作为推荐系统时，我们需要将原始数据(x_i, y_i)转换为带有提示模板的文本输入输出对(x_i^{text}, y_i^{text})。我们提供了一个模板示例，如图所示，其中包含了具体的转换流程*和方法。

Input:
The user is a male. His job is writer. His age is 35-45.
He watched the following movies in order in the past, and rated them:
['0. Pump Up the Volume (1990) (4 stars)', '1. Antz (1998) (4 stars)', "2. Devil's Own, The (1997) (5 stars)", '3. Crying Game, The (1992) (1 star)']
Based on the movies he has watched, deduce if he will like the movie ***Titanic (1997)*.**
Note that more stars the user rated the movie, the user like the movie more.
You should ONLY tell me yes or no.
Output:
No.

Figure 2: The illustration of textual input-output pair.

如图所示，文本输入$x_i^{文本}$包括用户的个人资料和历史行为，随后是一个关于对目标项目个性化偏好的分类问题。 为了从语言模型（LLMs）对应的关键答案单词（即Yes和No）的评分中获取浮点偏好估计$\hat{y}_i^{文本} \in [0,1]$，而不是离散的词令牌，我们通过双维softmax对相应的评分进行处理，在评估过程中完成偏好估计。

$$\hat{y}_i^{text} = \frac{\exp(s_{i,\text{Yes}})}{\exp(s_{i,\text{Yes}})+\exp(s_{i,\text{No}})} \in (0,1)$$

Low-Rank Adaptation of LLMs

低秩适应（LoRA）作为参数效率微调方法的热门选项，旨在减少大规模参数的LLMs进行微调时的资源消耗。因此，线性变换Y=XW被重新表述为：

$$Y' = XW + XAB^T = X(W + AB^T)$$

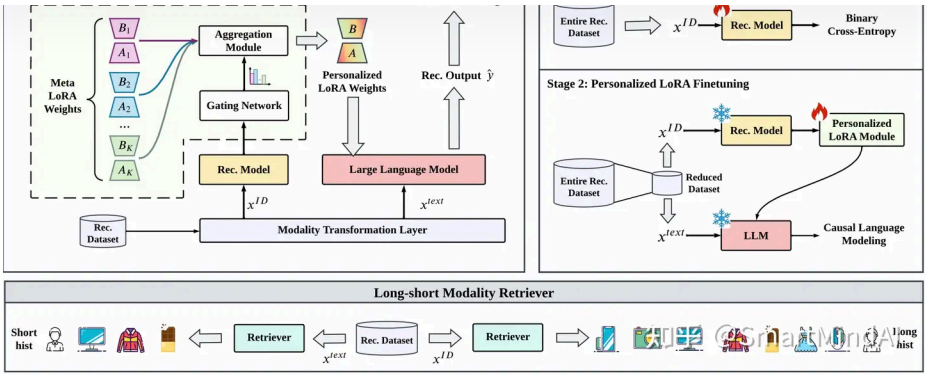
在其中，$X \in \mathbb{R}^{n \times d_{in}}$，$W \in \mathbb{R}^{d_{in} \times d_{out}}$，$A \in \mathbb{R}^{d_{in} \times r}$，$B \in \mathbb{R}^{d_{out} \times r}$，以及$r$远小于$min\{d_{in}, d_{out}\}$。

$$A \sim \mathcal{N}(0, \sigma^2),\ B = 0$$

为了使初始输出$Y' = XW + XAB^T$与$Y = XW$相等，在微调阶段W保持固定，而A和B通过基于SGD的优化方法如Adam进行更新。在使用LoRA对LLM进行微调时，我们有选择地将LoRA权重附加到LLM内部的特定类型的目标矩阵上。

Methodology

Overview of RecLoRA



我们提出RecLoRA框架来解决三个主要限制：

（1）为了在推荐中为语言模型（LLMs）构建个性化低秩适应，我们采用了一个包含个性化知识的并行meta-LoRA机制。

（2）为了减轻行为序列扩展相关的效果和效率问题，我们设计了一个长短期模态检索器。

（3）为了使语言模型的接受域扩展到整个训练空间，我们开发了一个Few2Many放大策略。在这一部分，我们将全面解释我们的RecLoRA框架，包括其网络架构和训练过程。

如图所示，RecLoRA接受两种输入组件：ID模态中的样本x^{ID}和文本模态中的样本x^{text}。每个模态x都有用户配置文件+、候选项目、上下文特征和用户历史行为，如式所规定。因此，我们首先将推荐数据集转换为这两种模态，然后分别将x^{ID}和x^{text}输入CRM和LLM模型。在转换过程中，长短期模态检索器为x^{ID}检索长期历史，为x^{text}检索短期历史。

$$
\begin{aligned}
x^{ID} &= LSMR(u^{ID}, v^{ID}), \\
x^{text} &= LSMR(u^{text}, v^{text}), \\
N_{u_i}^{ID} &> N_{u_i}^{text},
\end{aligned}
$$

其中N为序列长度+。一旦接收到x^{ID}和x^{text}，RecLoRA开始其处理过程。我们用于标记y的主要预测器是大型语言模型：

$$
y = LLM(x^{text}).
$$

因此，我们需要使用推荐数据集对预训练的LLM进行微调。为了实现内存效率，我们选择使用LoRA进行参数高效的微调，如第2.4{reference-type="ref" reference="lora"}节所述。然而，我们设计了一个个性化LoRA模块来整合用户个性化知识，而不是使用传统的LoRA。这个模块可以表示为：

$$
\begin{aligned}
h_{j+1} &= Layer_j(h_j) + LoRA(h_j), \\
h_{j+1} &= Layer_j(h_j) + PLoRA(h_j, x^{ID}),
\end{aligned}
$$

在LLM的第j层中$Layer_j$是我们模块的第j层，我们将其命名为PLoRA模块。该模块接受第j层的隐藏状态和x^{ID}作为输入，并输出个性化表示。

Personalized Low-Rank Adaption

针对个性化低秩适应，最直观的方法就是为每个用户u分别维护各自的LoRA权重(A_u, B_u)，也就是说，通过独立地为每个用户(A_u, B_u)维护LoRA权重，实现针对个性化低秩适应：

$$
Y_u' = XW + XA_uB_u^T, \ \forall u \in \mathcal{U}
$$

然而，这种方法的可扩展性较差，因为其模型复杂度随用户数量线性递增，导致内存消耗极其严重。此外，它忽略了不同用户行为模式之间的重叠和共性，可能导致性能不理想。 因此，为了平衡用户特定的个人偏好和用户之间的行为模式共性，我们提出了自定义LoRA模块。具体来说，如图所示，我们维护了一组元LoRA权重$\{(A_k, B_k)\}_{k=1}^{N_m}$，旨在捕捉多样的用户行为模式。每个元LoRA权重$(A_k, B_k)$可以视作隐式子空间，以增强表达能力和容量。 接下来，通过与可训练的门

知乎

$$A_p B_p^T = \sum_{k=1}^{N_m} \alpha_k \, A_k B_k^T$$

其中，A_p和B_p是最终的个性化LoRA矩阵，同时N_m代表元-LoRA权重的数量。在该公式中，第k个门控权重α_k表示当前用户和第k个LoRA权重的重要性及其相关性，用作个性化程度的指标。因此，为了捕捉复杂的相关性并实现个性化聚合，我们引入了传统的推荐模型（CRM）来为门控网络提供个性化的表示。CRM接收离散ID的输入，并生成隐藏状态作为个性化表示。
$R_c = CRM(x^{ID})$其中R_c是一个$n \times d_c$的实矩阵，代表CRM的输出。在获得表示之后，我们将其适应到闸门权重中，以引导元-LoRA权重根据个性化知识进行调整。我们使用具有ReLU激活函数[+]和层规范化（LayerNorm）的前馈网络（FFN）作为适配器：

$$\beta = W_2 \times \mathrm{ReLU}(\mathrm{LN}(W_1 R_c + b_1)) + b_2,$$

softmax 函数生成门控权重：

$$\alpha_j = \frac{\exp(\frac{\beta_j}{tp})}{\sum_{j'=1}^{N_m} \exp(\frac{\beta_{j'}}{tp})},$$

其中 $\alpha \in \mathbb{R}^{n \times N_m}$ 确保个性化 LoRA 权重的分布与元 LoRA 权重相匹配。

通过这种方式，此模块为每个用户动态生成一个唯一的个性化LoRA 模块，实现了个性化 LoRA 参数。

Few2Many Magnifier Strategy

在LLM微调过程中，理想情况下，应将整个推荐数据集暴露给LLM，LLM接受的数据量越多，其性能越好。然而，时间与效率的限制使得在输入端暴露大量数据变得不切实际。因此，一个更高效的方法是在不显著增加时间成本的同时，注入大量数据知识。为了实现这一目标，我们提出了"Few2Many放大策略"。

在第一阶段，我们使用完整的训练数据集来训练一个传统的推荐模型，以获得一个充分训练的模型。该模型能够在ID模态中，基于输入x^{ID}提供具有高泛化能力的样本级个性化向量表示，从而实现样本级别的个性化推荐。

$$\mathcal{L}^{ID} = \sum_{x_i^{ID} \in \mathcal{D}} [-y_i^{ID} \cdot \log(\sigma(\hat{y}_i^{ID})) - (1 - y_i^{ID}) \cdot \log(1 - \sigma(\hat{y}_i^{ID}))]$$

其中σ是Sigmoid函数[+]，y_i^{ID}是真实的标签$\hat{y}_i^{ID}$是在公式中计算的预测标签。现在，CRM已经拟合了完整的训练空间，并学到了整个数据集大部分的知识。

在第二阶段，我们对大规模训练集进行下采样[+]，以获得一个较小的训练集，用于RecLoRA的高效参数调整，其中包括大型语言模型。在这个过程中，第一阶段完全训练的传统推荐模型提供个性化信息，确保了空间视角的全面覆盖。这些信息被结合形成一个具有足够泛化能力的个性化LoRA矩阵。

因此，尽管大型语言模型在微调过程中只看到少量的训练样本，但它通过传统推荐模型和个性化LoRA矩阵扩展了其接收范围到完整的训练空间。这极大地提高了大型语言模型的样本效率和推荐性能。值得注意的是，在这个过程中，只有个性化LoRA模块被更新，而传统推荐模型和大型语言模型保持不变。用于训练的损失函数[+]是传统因果语言建模中的交叉熵损失[+]。

在推理阶段，大型语言模型不直接给出点对点评分$\hat{y}^{text} \in \{0, 1\}$因此，我们截取词汇评分，并对二进制关键答案词的评分进行双维度softmax操作。具体地，"是"和"否"的评分分别为s_y和s_n。然后，大型语言模型在推理阶段的点对点评分可以表示为：

$$\hat{y}^{text} = \frac{\exp(s_y)}{\exp(s_y) + \exp(s_n)}.$$

这个预测将用于计算评估指标。

Long-Short Modality Retriever

从当前样本预测中提取重要信息的有效方法是对长行为序列进行检索。然而，历史长度仍然是在有效性和效率之间做出权衡的关键因素。因此，我们在CRM中检索输入x^{ID}的长历史，在LLM中检索输入x^{text}的短历史。

CRM通过LoRA参数接受更长的历史信息，然后通过提供更多的序列信息和个性化知识来增强LLM。这种方法几乎不增加时间成本，实现了在有效性和效率之间非常出色的平衡，这一点将在第节中展示。

检索方法+可以多种多样且灵活。例如，我们使用语义行为编码来检索行为。具体来说，我们将每个行为输入到LLM中，并从LLM的最后层获得隐藏状态作为其表示。然后，我们应用主成分分析+（PCA）进行降维和去噪处理。最后，我们通过计算余弦相似度+来确定与当前项目最相关的前k个行为。

Experiment Setup

Datasets

我们对三个实际世界的数据集（即，GoodReads数据集，MovieLens-1M数据集和MovieLens-25M数据集）进行实验。

Table 1: The dataset statistics.

Dataset	#Users	#Items	#Samples	#Fields	#Features
MovieLens-25M	162,541	59,047	25,000,095	6	280,576
MovieLens-1M	6,040	3,706	970,009	10	16,944
GoodReads	449,114	1,432,348	20,122,040	15	5,787,895

Baseline Models

基于ID的传统特征交互模型，包含DeepFM、AutoInt、和DCNv2作为示例，以及GRU4Rec、Caser、SASRec、DIN和SIM作为用户行为模型的关键代表。另一方面，基于语言模型（LM-based）的类别中，包含CTR-BERT、TALLRec和ReLLa。

Implementation Details

Overall Performance (RQ1)

我们比较了RecLoRA与其他现有基线模型+的性能，并将结果报告在表中。在full-shot settings中，大多数其他推荐基线模型使用整个训练集进行训练。而TallRec、ReLLa和RecLoRA则在few-shot training sets上进行训练，这三个数据集的训练样本量均为70,000（大约占10%）。

Model		MovieLens-25M			MovieLens-1M			GoodReads		
		AUC	Log Loss	Rel.Impr	AUC	Log Loss	Rel.Impr	AUC	Log Loss	Rel.Impr
ID-based	DeepFM	0.8181	0.4863	3.52%	0.7978	0.5405	2.04%	0.7873	0.5027	1.47%
	AutoInt	0.8138	0.4942	3.77%	0.7976	0.5398	2.06%	0.7866	0.5024	1.56%
	DCNv2	0.8184	0.4905	3.50%	0.7977	0.5403	2.05%	0.7867	0.5022	1.55%
	GRU4Rec	0.8169	0.4917	3.55%	0.7959	0.5423	2.29%	0.7869	0.5020	1.52%
	Caser	0.8167	0.4917	3.39%	0.7952	0.5425	2.37%	0.7872	0.5023	1.49%
	SASRec	0.8163	0.4892	3.54%	0.7984	0.5388	1.97%	0.7864	0.5022	1.59%
	DIN	0.8258	0.4775	3.50%	0.7989	0.5389	1.90%	0.7880	0.4999	1.38%
	SIM	0.8379	0.4664	1.59%	0.7992	0.5387	1.86%	0.7896	0.4993	1.18%
LM-based	CTR-BERT	0.8079	0.5044	4.93%	0.7931	0.5457	2.64%	0.7176	0.5576	11.32%
	TallRec	0.8324	0.4733	1.65%	0.7789	0.5580	4.51%	0.7759	0.5148	2.96%
	ReLLa	0.8409	0.4662	1.22%	0.8005	0.5364	1.70%	0.7933	0.5055	1.93%
	RecLoRA	**0.8462***	**0.4552***	-	**0.8141***	**0.5248***	-	**0.7989***	**0.4916***	-

在表中，*Rel.Impr*表示RecLoRA相对于每个基线的相对AUC改进率。我们从表中得出的观察如下：

- RecLoRA在few-shot训练场景下表现出色，特别是在与TallRec、ReLLa和自身基线模型比较时。

知乎

- 在基于ID的所有基准模型中，SIM取得了最佳效果。通过采用用户行为检索，SIM有效地减少了用户序列中的噪音，这对于CTR预测非常有帮助。此外，传统的基于LM的CTR模型CTR-BERT的表现相较于大多数基于ID的传统CTR模型更差，这与[40, 65]中报告的结果一致。它仅整合了用于纯粹基于文本推荐的小型语言模型BERT[2]，从而导致性能不佳。
- ReLLa通常能够达到与最佳基于ID的基线模型SIM相媲美的性能。ReLLa整合了检索增强的指令调优（ReiT）技术，仅用了10%的训练数据。这个发现表明，对LLM进行微调以及在输入序列中结合检索的有效性。
- RecLoRA显著超越了最先进的基线，其显著性检验值小于0.01与最佳基线相比，这证明了我们提出的个性化LoRA模块的有效性。仅以10%的训练数据进行微调，LLM的性能得到了显著提升，这说明我们的RecLoRA模型能够精准地捕捉用户兴趣，从而实现更佳的用户个性化。

原文《Lifelong Personalized Low-Rank Adaptation of Large Language Models for Recommendation》

编辑于 2024-09-04 11:11 · IP 属地北京

LoRa 推荐系统 序列推荐

理性发言，友善互动

发布

还没有评论，发表第一个评论吧

推荐阅读

【LLM大模型面试】10个最常被问到的技术问题和答案（中...

你能解释一下Transformer架构及其在大型语言模型中的作用吗？ Transformer架构是一种深度神经网络架构，于2017年由Vaswani等人在他们的论文“Attention is All You Need”中首次提出。自那...

德国Viv... 发表于德国求职—...

LLM细节盘点(2)：位置编码

咸鱼王 发表于Arcit...

数据可视化Agent-企业应用中基于LLM的数据分析方案

一. 背景商业智能(Business Intelligence, BI)系统在企业商业决策中扮演着举足轻重的角色。在大型企业数据分析，预测等业务场景有着较大的应用场景。其中，ETL, 数据存储和BI智能分析，数据...

王大锤

为什么现在的LLM都是Decoder-only架构？从

本篇从理论、训练效率以及工现角度来阐述当前主流Deco only架构的优越性。 LLM是 “Large Language Model” 写，目前一般指百亿参数以上 言模型，主要面向 文本生成

喝拿铁的皮卡丘