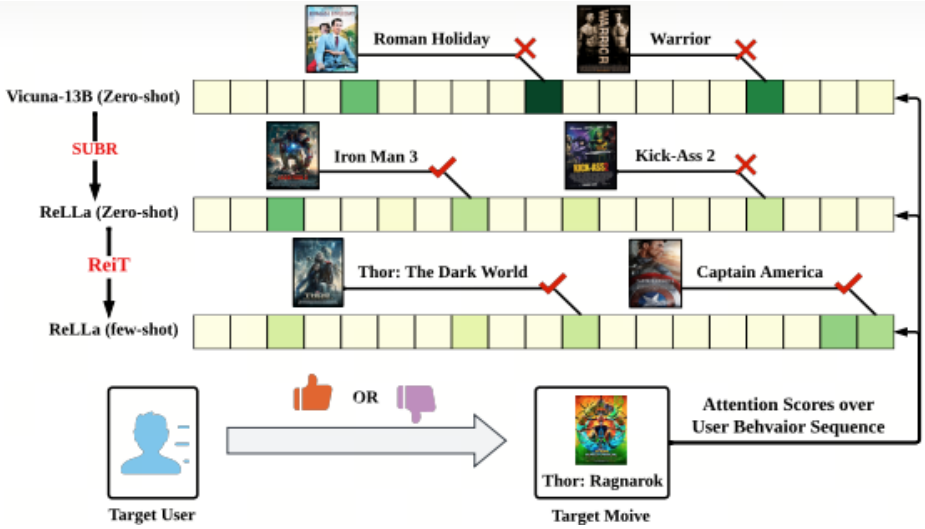
知乎



【华为2024】ReLLa：LLM在推荐系统中用户长序列行为的应用

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

LLMs因其在NLP中的强大能力，被研究者视为推荐系统的新潜力工具。它们能有效地处理推荐任务，包括零样本和少量样本，表现出强大的生成逼真文本的能力。近期的研究聚焦于将LLMs直接应用于推荐系统，如排序和评分，结果显示，无论样本数量多少，LLMs都能展现出优秀的性能。

本研究关注的是如何将大语言模型（LLMs）应用到零样本和少量样本的推荐系统，克服LLMs在理解和处理长用户行为序列时的挑战。具体地，我们提出"长期序列行为理解难题"，指出在面对长序列，LLMs未能有效提取推荐相关的信息，如在Vicuna-13B模型的2048个令牌限制下，性能随序列长度增加反而下降。传统推荐模型如SIM在序列增长时有所提升，但当行为数量超过窗口限制时，性能开始下滑。而在标准NLP任务中，LLMs在类似的大上下文环境下通常表现出色。

因此，我们在推荐领域面临的问题是如何让LLMs在有限上下文中推断用户对特定商品的偏好，这要求在用户档案和行为历史基础上进行复杂推理，这是LLMs独有的挑战。

Preliminaries

首先，收集用户的历史行为数据，比如浏览记录、搜索关键词或购买历史。然后，将这些行为转化为可以理解的语言表示，比如查询或者短语。接下来，将这些文本输入到LLM中，请求其生成相应的输出，如商品描述、评价或相关推荐结果。

利用LLMs进行点式评分通常涉及两个步骤：首先，LLM通过理解和生成文本，能够理解每个用户对特定商品或服务的潜在喜好。然后，当用户对某个产品提出查询或请求时，LLM可以根据其已有的知识库生成对该产品的评分或预测。这种评分可能基于用户的历史行为、商品属性或者LLM对相似情境的理解。由于LLMs的强大生成能力，可以生成高度相关的个性化评分，即使在缺乏直接评分数据的情况下。

Zero-shot and Few-shot Recommendations

(1) 零样本推荐是指直接运用预训练的LLMs，不依赖任何额外领域内特定数据，仅通过模型内在的通用知识和推理能力来为用户做推荐。

(2) 构建文本输入输出对时，首先获取用户信息和目标商品的相关信息，如用户历史行为和商品特性，转化为LLM能理解的格式，如查询或描述。然后，将这些输入送入模型，期望得到对商品的预测或评价。

习，生成对未知用户的精准预测。

Textual Input-Output Pair Formulation

对于LLMs，我们首先将每个样本x_i转化为文本格式的x_i^{text}，遵循特定的提示模板。同时，二元标签y_i以二进制形式

$$y_i^{text} \in \{\text{"Yes"}, \text{"No"}\}$$

进行转换。比如，一个输入可能包含用户对商品的描述、行为历史和任务说明。我们用图表展示了一个输入输出对(x_i^{text}, y_i^{text})的例子，其中x_i^{text}涵盖了所有相关的信息。

Input:
The user is a male. His job is college/grad student. His age is 25-34.
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: Illustration of textual input-output pair.

重要的是，用户行为序列的长度决定了生成上下文的范围，它可以从十几到几百不等。在处理每个样本x_i时，我们限制行为序列长度至K。比如在图中，实例使用$K = 4$。与之相比，一些标准的点击率预测通常会选取最近的K个行为，但ReLLa则独树一帜，通过语义用户行为检索，获取最相关于目标商品的那K个行为的文本信息，而非仅仅局限于这些行为本身。

Pointwise Scoring with LLMs

大型语言模型通过接受离散的文本tokensx_i^{text}，如$\mathbf{LanguageModel}$所示，生成对应的输出token$\hat{y}_i^{text}$。这个过程可以通过模型内部的训练函数来实现，该函数基于输入学习和推断，生成连贯且符合上下文的文本内容。用数学公式表示为：

$$\hat{y}_i^{text} = \mathbf{LanguageModel}(x_i^{text})$$

$$\begin{aligned} s_i &= \mathbf{LLM}(x_i^{text}) \in \mathbb{R}^V, \\ p_i &= \mathbf{Softmax}(s_i) \in \mathbb{R}^V, \\ \hat{y}_i^{text} &\sim p_i\,, \end{aligned}$$

在某些场景下，如点式评分或分类任务，LLMs可能会使用softmax函数来预测每个位置的输出。给定模型的隐状态s_i，softmax函数计算的是目标位置a和所有可能类别b（在这里是所有词汇V）的分数，通过指数函数加权，然后除以总和，得到一个概率分布：

$$\hat{y}_i = \mathbf{Softmax}_{a,b}(s_i) = \frac{e^{s_i(a)}}{\sum_{k \in V} e^{s_i(k)}}$$

这个概率向量表示了模型对各个类别（如"Yes"或"No"）的信念，常用于多分类问题中选择最可能的类别。在你的例子中$\hat{y}_i$就是用户对目标商品点击的概率估计。

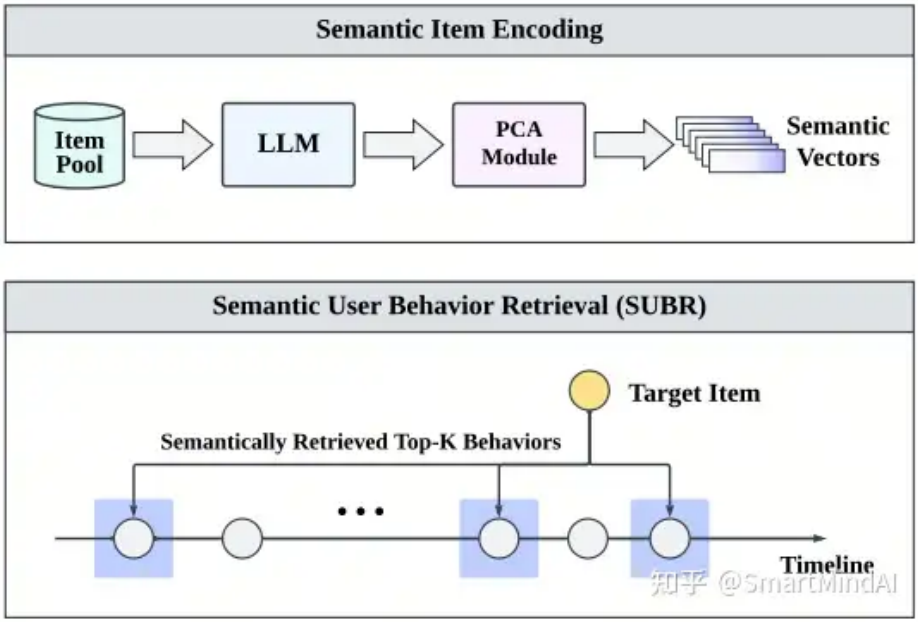
$$\hat{y}_i = \frac{\exp(s_{i,a})}{\exp(s_{i,a}) + \exp(s_{i,b})} \in (0, 1).$$

评估阶段使用的点击率估计$\hat{y}_i$是基于测试集的，这是为了量化模型对真实点击行为的预测能力。但在实际训练过程中，我们会结合LLMs的优化策略，遵循标准的指导原则，同时也会利用因果关系的语言建模方法，以确保模型的有效学习和优化。这样既能保持模型的性能，又能充分利用LLMs的潜在优势。

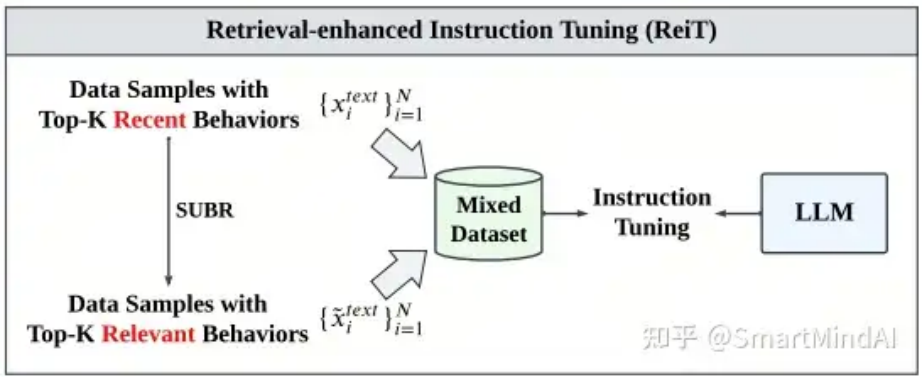
知乎

本文提出了ReLLa（Retrieval-enhanced Large Language Models，检索增强大型语言模型）框架。这个框架通过集成语义用户行为检索（SUBR）和检索增强的指令微调✦（ReiT），针对大型语言模型处理长用户行为序列的挑战，通过优化零样本推荐和少量样本推荐的任务，提升了模型的理解和性能。通过将数据样本x_i转化为文本输入x_i^{text}，并利用LLMs生成预测的输出$\hat{y}_i^{text}$，我们不仅进行了点式评分，还考虑了训练过程中的指导调优和因果语言建模的方法，确保模型的有效学习和应用。

Overview of ReLLa



在ReLLa框架中，我们专注于两种特殊推荐场景：零样本和少量样本。针对零样本，我们创新性地实施语义用户行为检索（SUBR）。首先，我们利用大型语言模型构建每个商品的语义向量。然后，对每个文本样本x_i^{text}，我们挑选与目标商品最相关的前K行为，取代原有的最近K行为，以提升样本质量。



对于少量样本推荐，我们设计了检索增强的指令微调（ReiT）。这种方法通过语义用户行为检索（SUBR）作为数据增强手段，将原始和检索增强的样本融合，丰富了训练数据，增强了模型在处理长序列行为时的理解和泛化能力。尽管ReLLa是在有限样本条件下优化的，但重要的是，与之对比，我们的基线模型✦是在全样本环境下进行训练的。

Semantic User Behavior Retrieval

在零样本情况下，由于缺乏领域内训练数据，我们采用语义用户行为检索（SUBR）来提升样本质量。具体来说，我们不是简单地使用最近的K个行为，而是找到与目标商品最相关的K个语义行为，用它们替换原始行为。这种方法旨在减少噪声，更精确地捕捉用户对目标商品的兴趣，同时保持行为序列的原始长度作为模型输入。

Figure 4: Illustration of descriptive text for an item (movie).

首先，我们进行语义项目编码，获取每个项目的语义表示。通过构造针对每个项目的描述性文本（如图所示），利用LLM处理并提取隐藏状态，平均得到一个维度为D（Vicuna-7B为4096，Vicuna-13B为5120）的向量$\boldsymbol{u_t} \in \mathbb{R}^D$。随后，通过主成分分析（PCA）降低维度并减少噪声，得到降维后的语义表示$\boldsymbol{v_t} \in \mathbb{R}^d$，其中$d = 512$。 接着，我们通过计算语义向量间的余弦相似性来评估它们的语义关联性。在零样本测试阶段，我们利用检索增强技术，用目标商品的最相关K个行为替换原始最近的K个行为，形成一个高质且保持上下文长度相似的并行增强测试集。这种做法显著提高了零样本推荐的性能，解决了处理长用户行为序列的难题。

Retrieval-enhanced Instruction Tuning

在少量样本推荐中，我们利用训练集

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

其中N代表样本总数。传统的做法是直接对LLMs进行转换后文本的指令微调，但我们注意到这可能导致过拟合和记忆衰退风险。因此，我们引入检索增强指令微调（ReiT），创新性地结合语义用户行为检索（SUBR）作为数据增强工具。 SUBR生成了多样化的用户行为模式混合训练集，通过增强每个样本$\tilde{x}_i^{text}$。我们合并原始和增强样本，形成包含$2N$个样本的混合训练集。这样的丰富模式有助于防止过拟合，提升模型在处理长序列行为时的泛化能力。 在微调过程中，我们遵循因果语言建模的原则，以保持模型结构的稳定，确保其有效学习和提取长期行为中的有用信息。

$$\max_\Theta \sum_{(x,y)\in\mathcal{M}} \sum_{j=1}^{|y|} \log P_\Theta(y_j|x, y_{<j})$$

在这个上下文中，Θ代表LLM的参数$\mathcal{M}$代表包含$2N$个混合训练样本来进行指令微调的集合。y_j是输出文本的第j个token$y_{<j}$指代前缀。我们采用二元交叉熵(BCE)进行CTR预测，仅在测试集上使用点式评分。

在微调过程中，我们使用SUBR增强数据以保持数据多样性，减少过拟合，增强模型的泛化能力。同时，通过这种方式，我们保持了对长序列行为的敏感性，同时保护用户隐私。

- ReiT通过SUBR进行的测试数据增强不会引入训练和测试数据的不一致性。这是因为SUBR基于语义相关性进行增强，保持了行为逻辑的连贯性，而非改变数据内容。这样的策略确保了增强数据与训练数据的目标一致性，有助于增强模型的泛化能力，而非引入额外的偏差。
- 一是通过增加两倍的训练样本来增加学习信息，这有助于减少过拟合；二是SUBR提供的模式丰富性作为正则化手段，它通过增加多样化的语义相关行为，帮助模型更好地理解和适应复杂用户行为，从而增强鲁棒性。实证研究表明，尽管两者都有积极影响，但模式丰富性在提升模型稳定性上起了决定性作用。

Experiment Setup

Datasets

我们对三个实际世界数据集，即BookCrossing、MovieLens-1M和MovieLens-25M进行了实验操作。这些数据集的信息如表所示。

Table 1: The dataset statistics.

Dataset	#Users	#Items	#Samples	#Fields	#Features
BookCrossing	278,858	271,375	17,714	10	912,279
MovieLens-1M	6,040	3,706	970,009	10	16,944
MovieLens-25M	162,541	59,047	25,000,095	6	280,576

Evaluation Metrics

为依据，包括AUC、Log Loss 和ACC 来评估方法的有效性。

Baseline Models

在点击率预测任务中，我们构建了两种主要基线模型：

(1) 传统CTR模型，包括(1) 特征交互模型（如DeepFM、AutoInt和DCNv2）和(2) 用户行为模型✦（如GRU4Rec、Caser、SASRec、DIN和SIM），这些模型通过处理一维离散ID；

(2) 基于语言模型的模型，如CTRBERT、PTab和P5，它们利用预训练的LM，处理用户行为的文本表示。 对于特征交互，我们使用平均池化方法提取用户历史行为作为附加特征。特别指出，我们加入了SIM模型，以确保比较的公正性，因为它引入了语义用户行为检索（SUBR）技术。

在实验中，我们不仅考察了AUC、Log Loss这样的标准度量，还关注了精确度分数（ACC）来全面评估模型✦性能。对于基于语言模型的基线，我们关注TALLRec，它通过简单指令微调对LLMs进行了优化，我们在后续的ablative研究中对此进行了深入探讨。

Implementation Details

我们选用FastChat的Vicuna-13B作为ReLLa的核心语言模型，实验环境为V100 GPU。为优化资源利用，我们应用8位量化和LoRA（低秩适应）进行参数高效的微调（PEFT），设定LoRA的秩为8，alpha为16，Dropout为0.05，仅作用于注意力块的查询和值投影。在指令微调中，我们采用AdamW优化器，无权重衰减，初始学习率分别为1×10^{-3}和1.5×10^{-3}，采用线性学习率调度。针对BookCrossing、MovieLens-1M和MovieLens-25M数据集，分别设最大训练轮数为10和5。在构建ReLLa的提示模板时，我们在处理BookCrossing数据集时，删除了'User ID'和'ISBN'，因为大型模型对这类纯标识符的处理效果不佳。类似地，对于MovieLens-1M，我们移除了'User ID'、'Movie ID'和'Zipcode'，而MovieLens-25M中去除了'User ID'和'Movie ID'。

Overall Performance (RQ1)

Model		BookCrossing				MovieLens-1M				MovieLens-25M			
		AUC	Log Loss	ACC	RelImpr	AUC	Log Loss	ACC	RelImpr	AUC	Log Loss	ACC	RelImpr
Zero-shot	Vicuna-7B	0.7011	0.9357	0.5378	3.45%	0.6739	0.9510	0.5644	4.07%	0.7468	0.6348	0.6392	-1.93%
	Vicuna-13B	0.7176	0.9507	0.5649	1.07%	0.6993	0.6291	0.6493	0.29%	**0.7503**	0.6308	0.6427	-2.39%
	ReLLa (Ours)	**0.7253***	**0.9277***	**0.5750***	-	**0.7013***	**0.6250***	**0.6507***	-	0.7324	**0.5858***	**0.7027***	-
Full-shot	DeepFM	0.7496	0.5953	0.6760	1.05%	0.7915	0.5484	0.7225	1.49%	0.8189	0.4867	0.7709	3.52%
	AutoInt	0.7481	0.6840	0.6365	1.26%	0.7929	0.5453	0.7226	1.31%	0.8169	0.4957	0.7689	3.77%
	DCNv2	0.7472	0.6816	0.6472	1.38%	0.7931	0.5464	0.7216	1.29%	0.8190	0.4989	0.7702	3.50%
	GRU4Rec	0.7479	0.5930	0.6777	1.28%	0.7926	0.5453	0.7225	1.35%	0.8186	0.4941	0.7700	3.55%
	Caser	0.7478	0.5990	0.6760	1.30%	0.7918	0.5464	0.7206	1.45%	0.8199	0.4865	0.7707	3.39%
	SASRec	0.7482	0.5934	**0.6811**	1.24%	0.7934	0.5460	0.7233	1.25%	0.8187	0.4956	0.7691	3.54%
	DIN	0.7477	0.6811	0.6557	1.31%	0.7962	0.5425	0.7252	0.89%	0.8190	0.4906	0.7716	3.50%
	SIM	0.7541	**0.5893**	0.6777	0.45%	0.7992	0.5387	0.7268	0.51%	0.8344	0.4724	0.7822	1.59%
	CTR-BERT	0.7448	0.5938	0.6704	1.71%	0.7931	0.5457	0.7233	1.29%	0.8079	0.5044	0.7511	4.93%
	PTab	0.7429	0.6154	0.6574	1.97%	0.7955	0.5428	0.7240	0.98%	0.8107	0.5022	0.7551	4.56%
	P5	0.7438	0.6128	0.6563	1.84%	0.7937	0.5478	0.7190	1.21%	0.8092	0.5030	0.7527	4.76%
Few-shot	ReLLa (<1%)	0.7482	0.6265	0.6800	-	0.7927	0.5475	0.7196	-	0.8335	0.4963	0.7779	-
	ReLLa (<10%)	**0.7575***	0.5919	0.6806	-	**0.8033***	**0.5362***	**0.7280***	-	**0.8477***	**0.4524***	**0.7925***	-

1. ReLLa在大规模数据集上的全面学习中表现出压倒性优势，无论在BookCrossing和MovieLens-1M上，它在AUC、Log Loss和Accuracy上明显优于Vicuna-13B，证明其能高效理解和处理用户行为数据。
2. 虽然在零样本测试的MovieLens-25M上，ReLLa的AUC指标稍有退步，但在点状指标如Log Loss和Accuracy上有所提升，这凸显了SUBR在降低大型模型解析难度上的价值，但也暗示了零样本学习可能带来的不稳定。
3. SUBR技术在全量学习中的成功应用，证实了它在减少模型对复杂用户行为的理解难度上的作用，即便在有限数据条件下也能有所改善。
4. 这种结果也提示我们，零样本学习在推荐任务中的应用并非始终稳定，尤其是在面对大数据集时，需要谨慎评估其适用性和稳定性。因此，ReLLa在不同学习模式下的表现，强调了策略选择的重要性。
5. SIM在所有基准模型中显示出最优性能，它通过用户行为的检索来滤除噪声，这对提高CTR预测的准确性至关重要。与之相反，基于语言模型的CTR模型，如CTRBERT、PTab和P5，尽管如文中所述使用了如BERT和T5这样的小型语言模型进行纯文本推荐，但在大多数基于ID的传统CTR模型中表现并不理想，这与之前的研究结果相符。这表明，仅仅依赖小规模语言模型可能导致他们在处理此类任务时表现不佳。

而相比之下，其他基线如SIM在MovieLens-25M上需要19,349,912个样本，这是整个训练集。这显示了ReLLa在序列推荐任务中的高效数据利用能力。

Sequential Behavior Comprehension (RQ2)

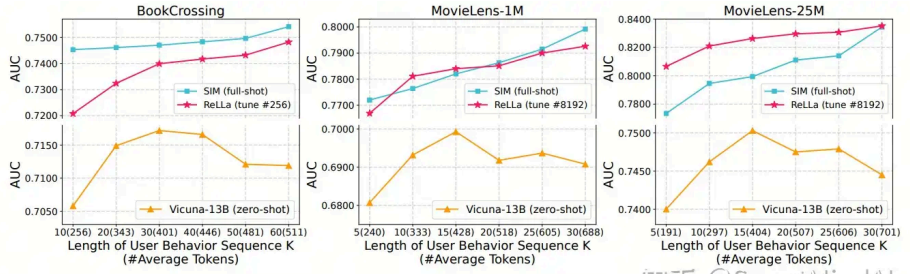


Figure 6: The AUC performance of different models w.r.t. different length of user behavior sequence K. ReLLa manages to mitigate the incomprehension problem of LLMs on recommendation tasks with long user behavior sequences.

1. 序列依赖性分析：随着观察窗口（K）增加，模型对时间序列+信息的捕捉能力增强，比如SIM和Vicuna-13B在较大K下性能通常提升，这体现了序列行为的连续性和复杂性。
2. ReLLa的优势展现：尽管SIM在全量学习下可能较为稳健，但ReLLa在多数情况下，特别是在合适K下，显示出在零样本学习上的优越性，这证实其在处理有限样本时的优越理解力。
3. 数据效率的体现：ReLLa在相对较少的数据下就能达到良好的效果，突显了其在序列推荐任务中的高效资源利用率。
4. 潜在挑战：然而，这也提出一个挑战，即零样本学习可能在处理大规模数据时不够稳定，需要权衡样本数量与性能之间的平衡。
5. 作为传统CTR预测模型，SIM在处理用户行为序列的全量学习版本（如K增大）时，其性能呈现出稳步增长的趋势，这符合我们通常对序列长度与信息深度关联的理解。更长的序列能为模型提供更丰富的行为历史，有利于提升推荐的有效性。
6. Vicuna-13B（零样本）在BookCrossing、MovieLens-1M和MovieLens-25M的实验中，其性能在序列长度$K = 30/15/15$时达到峰值，随后随长度增加而下降。值得注意的是，所用的令牌数量（约500/700/700）远未触及其2048个最大令牌限制。这表明在推荐任务的长序列处理上，即使大型模型如Vicuna-13B也面临理解和解析文本上下文的挑战，特别是需要对领域知识有深入理解。
7. ReLLa成功地克服了大型语言模型在处理长用户行为为序列时的理解挑战，相比之下，Vicuna-13B（零样本）在K大于30时在BookCrossing和K大于15时在两个MovieLens数据集上性能下滑。尽管如此，ReLLa在所有情况下都保持稳定，无明显性能下降，这与SIM相似，显示了它对长期序列的强大理解。随着K的增长，ReLLa的AUC评估指标持续提升，进一步确认了其在处理长序列时的有效性。

Data Efficiency (RQ3)

在探究样本效率时，我们通过调整样本数量N来考察ReLLa（少样本）和SIM（最优全量学习基线）。具体到实验N的取值分别为

$$128, 256, 512, 1024, 2048, 4096$$

（BookCrossing）以及

$$512, 1024, 2048, 4096, 8192, 65536$$

（MovieLens-1M和25M）。以$K = 60$（Book）、30（MovieLens-1M）和30（MovieLens-25M）设定用户行为序列长度。

知乎

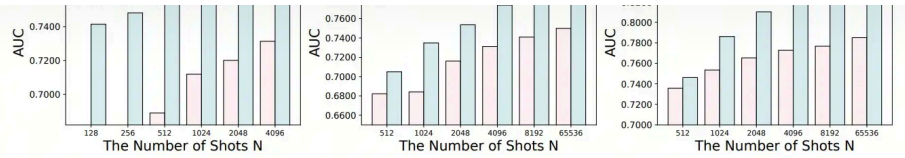
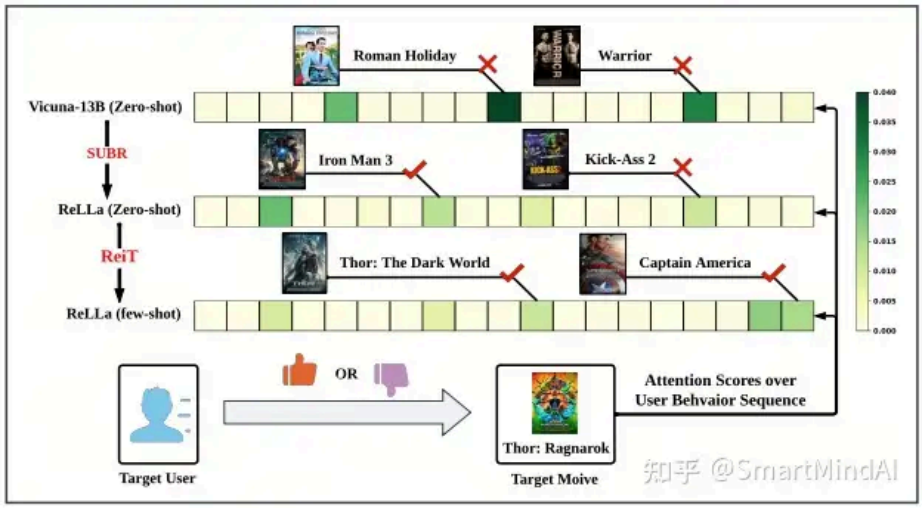


Figure 7: The AUC performance of ReLLa and SIM w.r.t. different numbers of shots N on three datasets, where "tune #N" indicates that we train the model with N training samples. The dashed line denotes the AUC performance of SIM (full-shot) that is trained with the whole training set. Notably, for $N = 128$ and $N = 256$ on BookCrossing dataset, few-shot SIM fails to accomplish the CTR prediction task, where the AUC is merely around 0.5, and is therefore omitted in the figure.

结果显示，无论是ReLLa还是SIM，随着N的增长，性能都有所提升。然而，ReLLa在所有N下都明显领先，且优势随N增加而更显著。特别在BookCrossing，当N低至128或256时，SIM在达到约0.5的CTR预估阈值上显得力不从心，这证明了ReLLa在有限样本下具有优秀的推理能力。这是因为ReLLa利用LLMs的深层逻辑理解和开放世界知识，展现出卓越的低样本数据+效率。

Case Study (RQ4)

我们通过实例研究考察了ReLLa如何助力大型语言模型（LLM）解析长用户行为序列。以MovieLens-25M数据集为例，我们对比了Vicuna-13B（零样本）、ReLLa（零样本）和ReLLa（少量样本）的表现。 Vicuna-13B在零样本环境下，对目标电影《雷神：诸神黄昏+》的关注主要分散在不相关影片如《罗马假日》和《勇士》，导致预测不精确。引入SUBR后，ReLLa（零样本）聚焦于与目标电影相似的超级英雄类型，如《钢铁侠3》，但仍存在不完全匹配的异常点，如《Kick-Ass 2》。 通过进一步的ReiT（检索增强的指令微调），ReLLa（少量样本）强化了对目标电影及其同属漫威系列的其他作品的注意力，这显示出SUBR和ReiT的协同作用。它们有效地帮助LLM理解了用户行为序列中的逻辑关系，从而提高了对目标电影的识别和预测准确性。总的来说，ReLLa通过优化样本质量和增强模型对长序列的处理，显著提升了LLM的性能。



原文《ReLLa: Retrieval-enhanced Large Language Models for Lifelong Sequential Behavior Comprehension in Recommendation》

发布于 2024-04-19 10:33 · IP 属地北京

推荐系统 LLM 序列

理性发言，友善互动