**How Can Recommender Systems Benefit from Large Language Models: A Survey**

**1.摘要**

研究团队对推荐系统如何从大语言模型中获益这一研究方向进行了全面的调研，并从现实世界中推荐系统的整个管道的角度鸟瞰。他们从两个正交的方面总结了现有的研究工作:在哪里以及如何将LLM适应于RS。对于“在哪里”的问题，他们讨论了LLM在推荐管道的不同阶段可以发挥的作用，即特征工程、特征编码器、评分/排名函数、用户交互和管道控制器。对于“HOW”问题，他们研究了训练和推理策略，得出了两个细粒度的分类标准，即是否在训练期间调优LLM，以及是否涉及常规推荐模型进行推理。对“WHERE”和“HOW”两个问题分别给出了详细的分析和一般的发展路径。然后，他们从三个方面强调了LLM适应RS的关键挑战，即效率，有效性和道德。最后，对调查结果进行了总结，并对未来进行了展望。

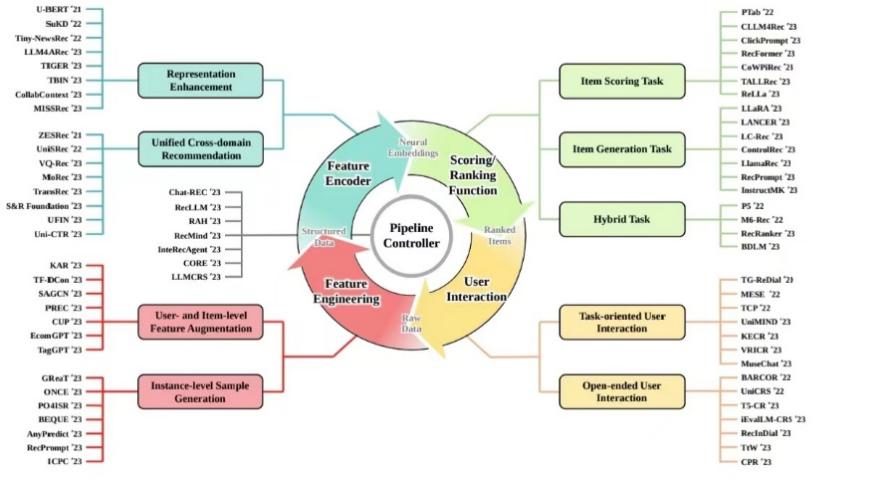
**2.引言**

随着在线服务的快速发展，推荐系统(RS)在匹配用户信息需求和缓解信息过载方面变得越来越重要，他们提供不同领域的个性化建议，它们估计给定用户对每个候选项目的偏好，并最终排列呈现给用户的项目排名列表。尽管传统的推荐系统在过去的几十年里取得了显著的进步，但它们的推荐性能仍然不是最优的，主要受到以下两个主要缺点的阻碍:

(1)传统的推荐系统是面向领域的系统，通常基于特定领域内的离散ID特征构建。因此，他们缺乏开放领域的世界知识来获得更好的推荐性能(例如，增强用户兴趣建模和项目内容理解)，以及跨不同领域和平台的转移能力。

(2)传统的推荐系统通常旨在以数据驱动的方式优化特定的用户反馈，如点击和购买，其中用户偏好和潜在动机通常基于在线收集的用户行为隐式建模。因此，这些系统可能缺乏推荐的可解释性，并且不能完全理解各种上下文中用户复杂而多变的意图。此外，用户无法通过提供详细的自然语言指令来主动引导推荐系统按照自己的需求进行定制推荐结果。

随着近年来大型基础模型的出现，它们在处理数据挖掘领域的许多具有挑战性的问题时提供了有希望的和通用的见解。一个代表性的形式是大型语言模型(LLM)，由于其对开放世界知识的巨大记忆，逻辑和常识推理能力以及对人类社会和文化的认识，它在各种语言处理任务中表现出了令人印象深刻的一般智能。通过将自然语言作为一种通用的信息载体，可以对不同形式、模式、领域和平台的知识进行综合、利用和解释。因此，我们是否可以结合LLM并从他们的共同知识中受益，以解决上述传统推荐系统根深蒂固的缺点。



**3.相关工作**

**3.1模型的评估**

得分/排名阶段的最终目标是高度与通用的推荐系统如2.1节中所讨论的,也就是说,提供一个排名项目列表(𝑖𝑘]𝑁𝑘= 1,𝑖𝑘∈我为目标用户𝑢∈U,我和你是普遍的项目集和用户(下一个项目的预测是一个特殊情况𝑁= 1)。当直接LLM得分/排名功能,适应这一目标可以通过各种任务LLM(例如,评级预测,生成项目ID)。根据LLM解决的任务不同，我们将相关研究工作分为三类:(1)项目评分任务，(2)项目生成任务，(3)混合任务。

3.3.1计分任务。在项目评分任务中，大型语言模型作为一个点函数(<e:1>)(𝑢，)，∀𝑢∈U，∀∈I)，它估计目标用户𝑢的每个候选项目的效用分数。这里U和I分别表示用户和项目的通用集。通过对候选集C中目标用户𝑢与每个条目的使用分数进行排序，得到最终的项目排序列表:C←prefilter(𝑢，I)，[𝑘]𝑘=1←Sort ({<s:1>(𝑢，)|∀∈I})，≤|C|，(2)，其中C是通过预过滤函数获得的候选集(例如，排序阶段的检索和预排序模型)。通过预滤波来减少候选项的数量，从而节省计算成本。预过滤器可以是推荐系统的第一个检索阶段的标识映射函数(即C = I)。

在不损失通用性的情况下，大型语言模型将文本提示符的离散标记符作为输入，并生成目标标记符𝑡-作为屏蔽语言建模中的屏蔽标记或因果语言建模中的下一个标记的输出。可以制定流程如下:ℎ= LLM(𝑥)𝑠= LM\_Head(ℎ)∈R𝑉𝑝= Softmax(𝑠)∈R𝑉ˆ𝑡∼𝑝,(3)ℎ最后表示,𝑉词汇量大小,𝑡ˆ是预测令牌从概率分布𝑝取样。

然而，项目评分任务要求模型对给定的用户-项目对(𝑢，)进行点对评分，并且输出应为实数(𝑢，)，而不是生成离散令牌𝑡。输出的结果要在一定的数值范围内，以表示用户的偏好。例如，对于CTR (click-through rate)的估计，要使用的结果中，泰瑞∈[0,1];对于评分的预测，要使用的结果中，泰瑞∈[0,5]。有三种主要方法可以解决这样的问题:输出需要连续的数值，而LLM生成离散的令牌。

第一类解[64,79,81,96,107,115,197,199,260,271,274]采用单塔范式[155,256]。

具体来说，他们直接放弃了语言建模解码器头(即LM\_Head(·))，将Eq. 3中LLM的最终表示<e:1>输入到一个精心设计的投影层中，计算分类或回归任务的最终分数，即:=(𝑢，)= MLP()，(4)。

其中MLP (multi-layer perceptron的缩写)是投影层。输入提示符≥需要同时包含来自用户𝑢和条目<e:1>的信息，以支持基于单个潜在表示<e:1>的偏好估计。

CoLLM[260]和E4SRec[96]利用预先学习的用户和项目ID嵌入构建个性化提示，以精确估计偏好。FLIP[199]和ClickPrompt[107]分别提出以并行和堆叠的方式对语义信息和协作信息进行细粒度的知识对齐和融合。

CER[157]加强了推荐与其自然语言解释之间的一致性，以提高评级预测性能。Kang等[79]以回归的方式对大型语言模型进行了微调，通过将微调后的LLM模型大小缩放到110亿，显示出令人惊讶的性能。该研究领域的其他典型例子包括:LSAT[174]、BERT4CTR[197]、CLLM4Rec[271]、PTab[115]。

与第一种方法相似，第二种解[87,128,188,190,191,234]也抛弃了LLM的解码器头。然而，它的与众不同之处在于它采用了传统推荐系统中流行的双塔结构[53,54,209]。它们分别维护两个独立的塔来获得用户和项目的表示，并且通过两个表示之间的一定距离度量来计算偏好得分:=(𝑢，)=𝑑(𝑇𝑢(𝑢)，𝑇())，(5)，其中𝑑(·，·)是距离度量函数(例如余弦相似度，L2距离)。𝑇𝑢(·)和𝑇萨拉赫(·)是由LLM主干组成的用户和条目塔，用于从用户和条目文本(即，≥𝑢和≥≥)中提取有用的知识表示。在这一系列的工作中，设计了不同的辅助结构来增强LLM的双面信息。例如，CoWPiRec[234]将词图神经网络应用于用户行为序列中的项目文本，以扩大语义信息的相关性。通过使用编码器-解码器LLM, TASTE[128]首先将每个用户行为编码为软提示向量，然后利用解码器从软提示序列中提取用户偏好。其他典型例子包括:RecFormer[87]、LLM-Rec[188]、CUP[191]。

上述两种解决方案都是用人工设计的预测模块替换原始语言建模解码器头(即LM\_Head(·)解决方案[3,56,57,111,130,135,149,156,176,182,223,226,259,261,265,273]则提出保留解码器头，并基于概率分布𝑝∈R - m进行偏好估计。TALLRec[3]、ReLLa[111]、PromptRec[226]、BTRec[57]和CR-SoRec[149]在用户个人资料、用户行为和目标项目的文本描述之后，对用户偏好追加一个二进制问题，从而将项目评分任务转化为二进制问答问题。然后,他们可以拦截估计分数𝑠∈R𝑉或者概率𝑝∈R𝑉Eq。3,进行一个二维的softmax超过相应的分对数的二进制关键答案的话(例如,令牌用来表示标签,例如,Yes / No)得分点态:ˆ𝑦= exp(𝑝𝑌𝑒𝑠)exp(𝑝𝑌𝑒𝑠+ exp(𝑝𝑁𝑜)∈(0,1),(6)𝑝𝑌𝑒𝑠和𝑝𝑁𝑜表示“是”和“不”的分对数令牌,分别。提取项目评分的相应标签令牌的softmax概率的其他典型示例包括TabLLM [56]， Prompt4NR[261]和GLRec[223]。此外，另一项研究打算将项目描述(如标题)与不同模板的用户行为历史连接起来，并通过计算提示文本的总体困惑度[135,156]、对数似然度[171,176]或联合概率[259]来估计分数，作为用户偏好的最终预测分数<e:1>。

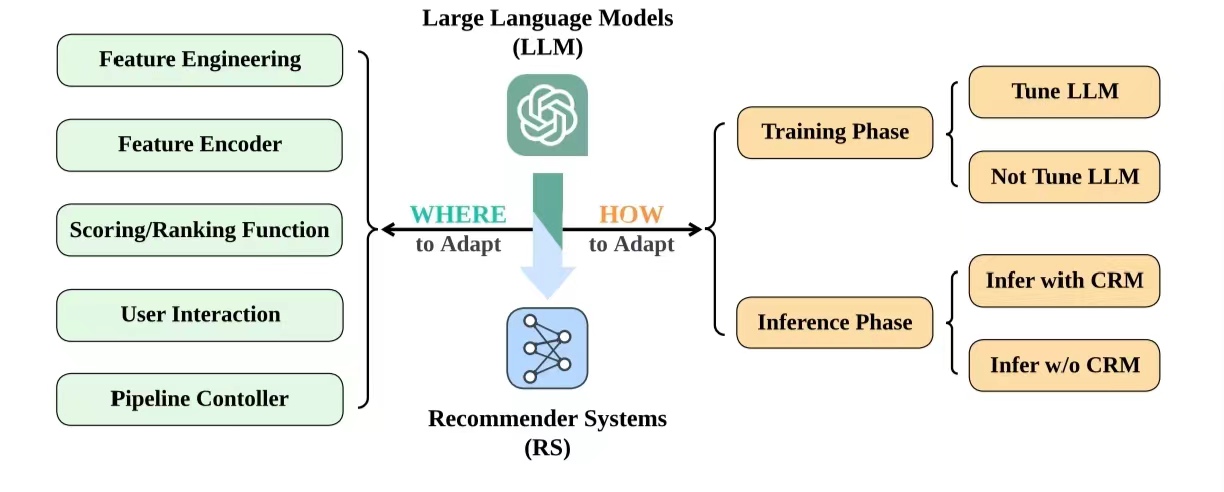
此外，Zhiyuli等[265]指导LLM以文本方式预测用户评分，并通过人工设计提示符将输出格式限制为小数点后两位的值。

3.3.2项目生成任务。在项目生成任务中，大语言模型作为生成函数(𝑢)直接生成项目的最终排名列表，只需要一次前推函数(𝑢)。一般来说，物品生成任务高度依赖于LLM的内在推理能力来推断用户偏好并生成排序物品列表，其过程可以表述为:[𝑘]𝑘=1 =(𝑢)，𝑠。𝑡。(7)根据是否为LLM提供了一组候选项目来完成项目生成任务，我们可以将相关的解决方案分为两类:(1)开集项目生成，(2)闭集项目生成。

在开集项目生成任务[2、30、45、51、65、69、73、88、89、98、105、112、137、153、170、238、249、263、270]中，LLM需要在没有给定候选项目集的情况下，根据用户简介和行为历史直接生成用户可能喜欢的排序项目列表。由于在输入提示符中没有提供候选项目，因此大型语言模型实际上并不知道通用项目池I，从而带来了生成幻觉问题[137]，生成的项目可能无法与项目池I中的确切项目匹配。因此，除了输入提示符模板的设计[62,100]和微调算法[89]外，为了克服生成幻觉问题，还需要在生成后进行项目接地和匹配的后处理操作[137]。我们制定过程如下:ˆ𝑖𝑘𝑁𝑘= 1←LLM(𝑥𝑢),[𝑖𝑘]𝑁𝑘= 1←匹配ˆ𝑖𝑘𝑁𝑘= 1,我,(8)匹配(··)是匹配函数,ˆ𝑖𝑘LLM-generated物品,和𝑖𝑘是我根据的实际项目匹配ˆ𝑖𝑘。LANCER[73]采用知识增强的前缀调优作为生成基础，并进一步应用余弦相似度将生成的项目文本的编码表示与通用项目池进行匹配。Di Palma等人[30]利用ChatGPT进行用户兴趣建模，并利用Damerau-Levenshtein距离[138]生成下一个项目的标题进行项目匹配。

除了以文本方式生成项目外，另一项研究侧重于将语言空间与基于id的推荐空间对齐，从而使LLM能够直接生成项目id。例如，Hua等人[65]探索了更好的项目索引方法(例如，顺序索引、协作索引)，以提高这类索引生成任务的性能。LightLM[137]设计了一个轻量级的LLM，精心设计了用户和项目索引，并应用约束束搜索来生成开集项目ID。此外，LLaRA[105]使用一种新颖的混合方法来表示LLM输入提示中的项目，该方法将传统推荐器中基于id的项目嵌入与文本项目特征集成在一起。其他典型的开集项目生成方法包括:GenRec[71]、TransRec[112]、LC-Rec[263]、ControlRec[153]和POD[89]。

在闭集项目生成任务[16,45,46,62,113,130,133,178,185,196,206,214,230,236,243,251]中，LLM需要对给定的候选项目集进行排序或选择。也就是说，我们将首先采用一个轻量级检索模型，将通用项目集I预过滤为有限数量的候选项目，表示为C = {<e:1>𝑗}𝐽𝑗=1，𝐽≪|I|。由于LLM的上下文窗口限制，候选项目的数量通常设置为20个。然后将候选项的内容呈现在LLM的输入提示符中，以生成排序的项列表，该列表可以表示为:C←预过滤(𝑢，I)，[𝑘]预置𝑘=1←LLM(𝑢，C)，预置≤|C|， (9)



**3.2在哪里适应大语言模型**

我们通过详细阐述LLM对推荐管道不同部分的适应来回答“在哪里”的问题(1)特征工程，(2)特征编码器，(3)评分/排名函数，(4)用户交互，(5)管道控制器。值得注意的是，由于LLM的多任务性质，在同一项研究工作中使用LLM可能会涉及到推荐管道的多个阶段。例如，LLM在CUP的特征工程和评分/排名功能两个阶段都得到了利用。

**3.3如何适应大语言模型**

为了回答如何使LLM适应RS的问题，我们采用了两个正交的分类标准来区分LLM对RS的适应，得到了如图4所示的四象限分类:•Tune/Not Tune LLM表示我们是否会在训练阶段根据域内推荐数据对LLM进行调优。调优LLM的定义既包括全调优，也包括其他参数高效的调优方法(如LoRA[63]、提示调优[82])。

•有/没有CRM的Infer表示我们是否会在推理阶段涉及传统的推荐模型(CRM)。请注意，有些作品只使用CRM作为独立的预过滤函数来生成LLM的候选项目集[46,196,243]。我们将它们归类为“没有CRM的推断”，因为CRM独立于LLM，并且可以与最终的推荐任务解耦。

在图4中，我们使用不同的标记大小来表示研究工作所适应的大型语言模型的大小，并使用不同的颜色来表示他们在项目推荐方面击败的最佳基线。因此，由于没有提供传统的推荐评估，少数作品没有呈现在图4中，例如，RecLLM[39]只研究了涉及RS管道控制的LLM的系统架构设计，而没有进行实验评估。

此外，值得注意的是，一些研究工作可能会提出跨不同象限应用的技术。例如，ReLLa设计了语义用户行为检索，以帮助LLM在零次预测(即象限3)和少次微调(即象限4)设置中更好地理解和建模终身用户行为序列。

根据四象限分类法，我们证明了“HOW”研究问题的整体发展路径通常遵循图4中的浅色箭头。因此，我们将按照象限1,3,2,4的顺序介绍最新的研究成果，然后对每个象限分段进行详细的讨论。

**3.4面对的挑战**

在本节中，我们强调了使LLM适应RS的主要挑战，这些挑战主要来自于推荐系统和现实世界应用程序的独特特征。因此，我们也将讨论现有工作的初步努力，以及其他可能的解决方案。从三个方面提出了以下挑战:(1)效率(训练效率，推理延迟)，(2)有效性(域内长文本建模，ID索引和建模)，(3)伦理(公平性，以及LLM的其他潜在风险)。

**4.方法**

**4.1.1用CRM调整LLM & Infer(象限1)**

象限1是指在训练阶段对具有域内推荐数据的大型语言模型进行微调，同时引入常规推荐模型以提供更好的协同知识的研究工作。

LLM(如ChatGPT)的兴起展示了令人印象深刻的新兴能力，如推理和指令遵循，为LLM增强推荐指出了有希望的方向。研究人员开始研究将十亿级大型语言模型(如LLaMA[192]和ChatGLM[36,248])纳入推荐系统领域的潜力。

•由于LLM拥有大量的模型参数，我们很难对LLM进行完全的调优，因为它会导致计算资源的不可承受的成本。为了提高训练效率，通常采用参数高效微调(PEFT)方法，通常需要更新的参数少于1%。

•LLM的角色不再是CRM的简单可调功能编码器。为了更好地利用LLM所展示的推理能力和开放世界知识，研究者倾向于将LLM和CRM置于同等地位(例如，两者都是推荐者)，相互利用各自的优势进行协作，从而提高推荐性能。

由于涉及到CRM，并且LLM是可调的，因此象限1的研究工作可以更好地与推荐系统的数据分布保持一致，因此即使所涉及的LLM规模相对较小，也都可以获得令人满意的性能。我们可以观察到一个明显的趋势，即研究人员打算考虑从百万级到十亿级的更大的语言模型，从而受益于它们大量的开放世界语义知识，以及指令跟随和推理能力。

**4.1.2在没有CRM的情况下不调整LLM和Infer(象限3)**

象限3是指排除常规推荐模型，仅采用冻结的大语言模型作为推荐者的研究工作。这类研究通常是在大型基础模型出现后出现的，尤其是ChatGPT，研究人员的目标是分析LLM在推荐领域的零概率或少概率性能，其中LLM是固定的，排除了CRM。

ReLLa[111]提出执行语义用户行为检索，将简单截断的top-𝐾最近行为替换为与目标项目top-𝐾语义相关的行为。这样可以提高数据样本的质量，从而使LLM更容易理解用户序列，获得更好的零采样推荐性能。RecMind[213]设计了自我激励的提示策略，使LLM能够通过额外的工具显式访问外部知识，例如SQL用于推荐数据库，搜索引擎用于web信息。Chat-REC[43]指示ChatGPT不仅作为评分/排名函数，而且还控制推荐管道，例如决定何时调用独立的预排名模型API。

LLM所包含的知识是全局的、事实性的，而推荐是一项个性化的任务，需要面向偏好的知识。这说明了来自推荐系统训练数据的领域内协作知识的重要性，目前仅仅依靠固定的大型语言模型是不适合很好地解决推荐任务的。

因此，进一步为LLM注入域内协作知识以提高推荐性能主要有两种方法:(1)涉及CRM进行推理，(2)基于训练数据对LLM进行调优，分别参考象限2和象限4的工作。

**4.1.3未调整LLM和Infer与CRM(象限2)**

象限2的研究工作利用LLM的不同关键能力(如丰富的语义信息、推理能力)，未经微调，帮助CRM更好地完成推荐任务。与象限1中的工作类似，象限2中冻结LLM的使用通常展示了LLM随时间演进的大小的发展路径，即从小规模语言模型到大型语言模型。

随着模型规模的扩大，涌现能力和丰富的开放世界知识使大型基础模型能够将其作用扩展到推荐管道的其他阶段，如特征工程阶段[85,119,215,228]和用户交互阶段[55,61,165,210]。AnyPredict[215]利用ChatGPT api来整合表格样本，以克服具有不同模式的表之间的障碍，从而为后续的传统预测模型提供统一的扩展训练数据。ONCE[119]利用ChatGPT进行新闻片段生成、用户分析和新闻摘要，从而用llm生成的特征增强了新闻推荐模型。KAR[228]和RLMRec[166]利用LLM通过特别设计的输入模板和链式提示策略来增强用户行为建模，旨在为CRM提供用户级功能增强。Wang等[210]和He等[55]研究了整合LLM(如ChatGPT和GPT4)来处理会话推荐用户交互过程中的开放式自由形式聊天。

在这些作品中，虽然LLM是冻结的，但与象限3(即Not Tune LLM;根据他们击败的最佳基线来推断(没有CRM)。与象限1(即Tune LLM;由于大型语言模型是固定的，象限2中的LLM的作用主要是在推荐管道的不同阶段辅助CRM，包括但不限于特征工程和特征编码器。

**4.1.4在没有CRM的情况下调整LLM和Infer(象限4)**

象限4的研究工作旨在微调大型语言模型，使其作为评分/排名功能，基于推荐系统的训练数据，排除CRM的参与。由于排除了CRM，我们必须应用提示模板来获得文本输入输出对，并因此将推荐任务(例如，点击率估计和下一项预测)转换为文本分类任务[3,111]或序列到序列任务[44,178,196]。

通常需要参数有效的微调方法(PEFT)来有效地使十亿级LLM适应RS，其中低秩自适应方法(LoRA)是最受欢迎的选择。例如，ReLLa[111]、GenRec[70]、BIGRec[2]、RecSysLLM[21]和LSAT[174]采用LoRA[63]技术对基本大型语言模型(通常为LLaMA-7b[192]或Vicuna-7B[18])进行项目评分或生成任务的调优。

除了LoRA, M6- rec[23]还将选项调优设计为提示调优的改进版本，使M6[109]能够执行各种下游任务，如项目检索和排名。VIP5[45]通过推荐任务的通用基础模型进行分层适配器调优，统一各种模式(如ID、文本和图像)。

尽管PEFT的引入缓解了训练效率低下的问题，但对于训练样本数量可能达到数十亿的实际应用程序来说，计算开销仍然过高。在这种情况下，即使是像LoRA这样的PEFT方法也不足以让LLM遍历整个训练数据集。为此，最近的研究开始研究LLM的强归纳学习能力[150]，方法是将整个训练集降采样为一个小规模子集[16,79,11,129]。作为代表，ReLLa[111]统一采样不到10%的训练实例，并令人惊讶地发现，仅基于不到10%的样本进行微调的LLM能够优于在整个训练上训练的传统推荐基线模型

**4.2.1效率和推理延迟方面的调优方法**

提高现代基于深度学习的推荐系统的性能有两个关键方面:

1. 扩大训练数据量(例如十亿级别的训练样本)
2. 提高模型的更新频率(从日级到小时级，甚至分钟级)。这两个因素都对训练效率提出了很高的要求。

我们可以减少LLM的训练数据量并放宽更新频率(例如，周级)，同时保持CRM的完整训练数据和高更新频率。支持这种方法的基础是研究人员[16,11,266]指出LLM具有很强的归纳学习能力，可以通过少量监督产生广义和可靠的输出。通过这种方式，LLM可以为CRM提供一致的领域内知识，而CRM则充当LLM的频繁更新适配器。

在线推荐系统通常是实时服务，对时间非常敏感，所有阶段(如匹配、排名、重新排名)都应该在几十毫秒内完成。LLM在推理阶段的参与导致了推理延迟问题。LLM的推理时间相当高，更不用说快速生成模板带来的额外时间成本。

当我们必须在推理阶段涉及到LLM时，预计算和缓存LLM的输出或中间表示可以作为确保低延迟推理的常用策略。在采用LLM作为评分/排序函数时，M6-Rec提出了多段后期交互策略。用户和项目的文本特征被分割成更静态的细粒度片段，例如，通过将每个点击的项目表示为一个单独的片段。然后，我们可以使用前几个转换层预先计算和缓存每个段的编码表示，而其余层则用于在推荐请求到达时执行多个段之间的后期交互。其他作品如UniSRec和VQ-Rec只是采用语言模型作为特征编码器。因此直接缓存由22生成的密集嵌入是很简单的。语言模型。预计算和缓存策略可能适合于项目端信息，因为它们通常是静态的，但对于用户端信息可能不是最优的，因为用户行为和兴趣是高度动态的，并且随着时间的推移会迅速发展。因此，我们必须找到一个合适的缓存频率来平衡性能和计算成本。

我们还可以寻求减少模型大小以提高推理效率的方法，这些方法已经在其他深度学习领域得到了很好的探索，例如蒸馏[74]，修剪[15]和量化[246]。例如，CTRL[95]和FLIP[199]提出通过对比学习将LLM的语义知识提炼到CRM。然后使用改进的参数初始化对CRM进行单独调优，以获得更好的推荐性能，同时保持低延迟推理。这些策略通常涉及模型性能和推理延迟之间的权衡。或者，我们可以在特征工程阶段引入LLM，并预先存储LLM的输出，这将为推理带来更小(但并非完全可以忽略不计)的额外负担。

我们还可以将LLM引入到会话推荐系统等推理延迟约束相对宽松的场景中。

**4.2.2长文本建模方面的调优方法**

在NLP领域，提出了一系列的工作来解决上下文窗口限制(例如，滑动窗口[216]，记忆机制[32])，这可以在推荐系统中考虑。最近的研究提出结合CRM的潜在表示来压缩LLM的个性化输入提示，从而缓解长文本问题。例如，CoLLM[260]和E4SRec[96]通过线性投影层将每个用户行为的文本描述替换为CRM嵌入表中的一个潜在向量映射，这大大减少了长用户序列的令牌数量。受前缀调优[101,123]的启发，ClickPrompt[107]将CRM的样本最终表示转换为LLM的分层提示，从而更容易从提示模板中消除不必要的功能。

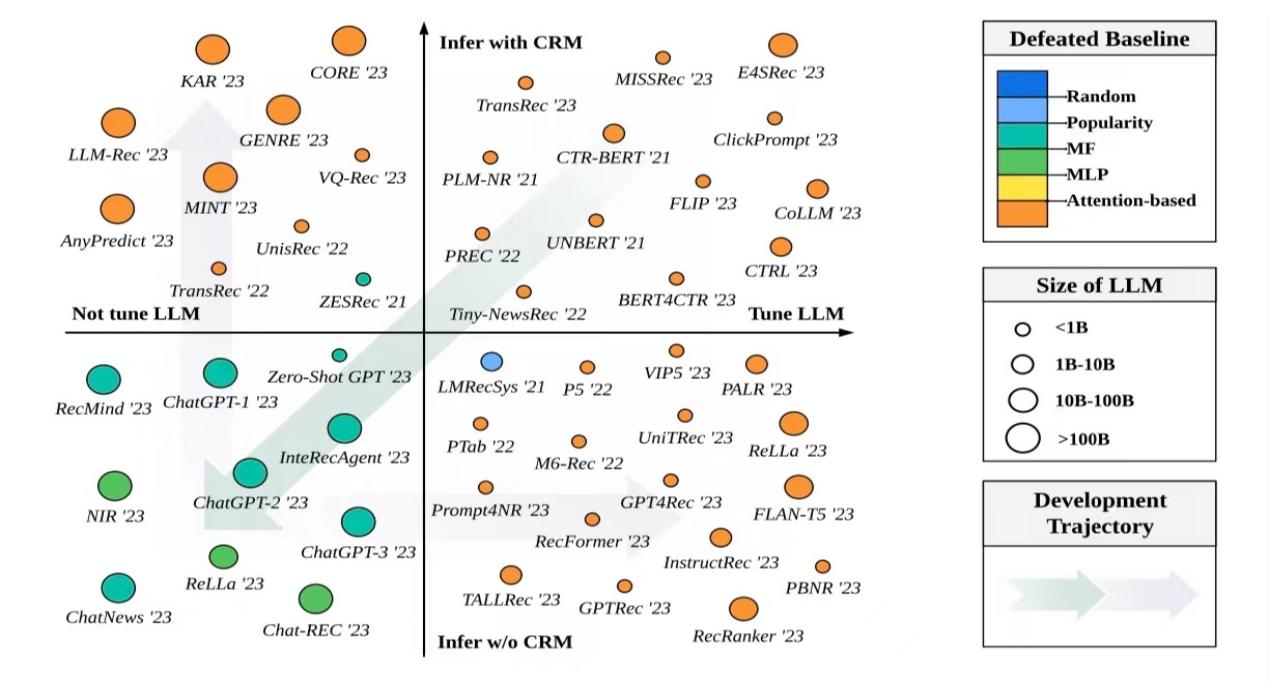
**5.实验**

**一．**

我们可以观察到，LLM如何适应RS的发展路径与大型语言模型的进展是基本一致的。在2021年和2022年初，预训练语言模型的参数大小还比较小(例如BERT-base为110M, GPT2-XL为1.5B)。因此，早期的工作通常倾向于将小规模的语言模型作为文本特征编码器，或者作为评分/排名函数进行微调以适应推荐系统的数据分布。推荐过程被简单地表述为一个一次性的直接预测任务，并且可以在语言模型的帮助下更好地解决。

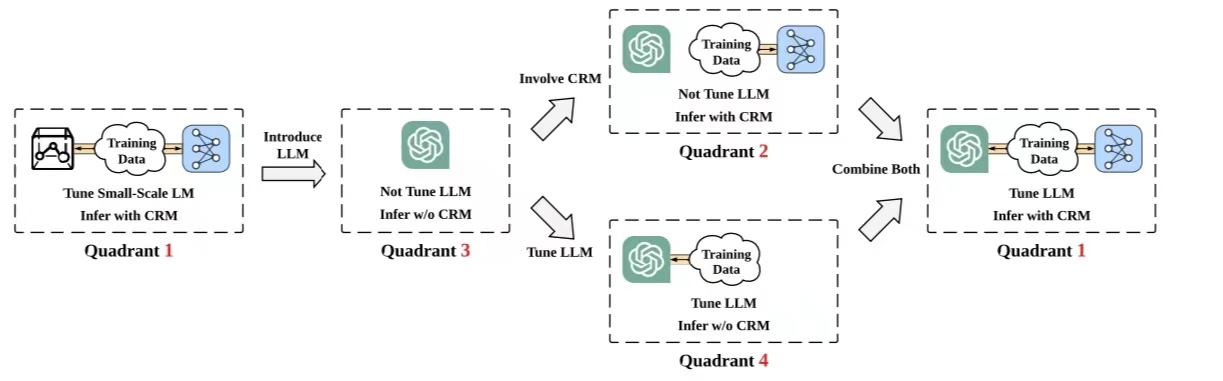
随着模型尺寸的逐渐增大，大型语言模型获得了突现能力(如指令跟随和推理)，以及大量具有强大文本生成能力的开放世界知识。有了这些大规模参数带来的惊人特性，LLM不仅开始深化其在特征编码器和评分/排序函数阶段的使用，而且将其作用进一步扩展到推荐管道的其他阶段。例如，在特征工程阶段，我们可以指导LLM生成可靠的辅助特征和合成数据样本，以辅助模型的训练和评估。通过这种方式，将LLM的开放世界知识注入到闭域推荐模型中。此外，大型语言模型还为各种信息系统提供了更人性化的自然语言界面和自由形式的对话，从而彻底改变了用户交互。。

总之，随着大型语言模型的能力得到进一步的探索，它们将与推荐管道的多个阶段逐渐形成更深层次的耦合和绑定。更进一步，我们可能需要定制专门定制的大型语言模型来满足推荐系统的独特需求。



我们首先总结了LLM适应RS时协同知识注入的必要性，然后从“如何”的问题上总结了整体的发展路径，以及未来可能的方向。接下来，我们讨论了推荐性能与适应的LLM大小之间的关系。最后，我们讨论了关于大型语言模型的硬样本重排序的一个有趣的性质。

**二．**



**1.需要知识协同。**从图中，我们可以看到象限3和象限1,2,4的作品之间有清晰的性能边界。象限3的研究工作即使采用大规模模型(即ChatGPT或GPT4)，即使配备了用户行为检索和工具使用等先进技术，也是低劣的。这表明推荐系统是一个高度专业化的领域，需要大量的领域内协作知识。LLM无法从它的通用预训练语料库中有效地学习这些知识。因此，在使LLM适应RS时，我们必须涉及领域内的协作知识以获得更好的性能，通常有两种方法来实现这一目标

•在训练阶段调整LLM，从以数据为中心的方面注入协作知识。

•在推理阶段使用CRM进行推断，从以模型为中心的方面注入协作知识。

当llm适应RS时，这两种方法都强调了领域内协作知识的重要性。

基于以上的认识，如图5所示，我们在四象限分类的基础上，绘制了关于“HOW”研究问题的总体发展趋势。从2021年初开始，研究人员通常打算将小规模LM和CRM结合起来，对推荐进行联合优化(即象限1)。

然后，在2023年初左右，一些作品开始在没有CRM帮助的情况下利用冻结的LLM进行推荐(即象限3)，其较差的表现表明了协作知识的必要性。为此，本文提出了两种主要的解决方案，分别对应象限2和象限4，通过涉及CRM或调优LLM进行域内协同知识注入。接下来，当我们发现LLM适应RS的黄金原则(即领域内协作知识注入)时，发展路径进一步回到象限1，我们的目标是共同优化LLM和CRM以获得卓越的推荐性能。最后，在如何使LLM适应RS方面，未来可能的方向可能是如何将来自推荐系统的协作知识与LLM所展示的通用语义知识和应急能力更好地结合起来。例如，为基于代理的LLM提供外部工具，以便更彻底地访问推荐数据，以及来自搜索引擎的实时web信息。

**2.越大越好吗?**通过注入以数据为中心或以模型为中心的领域内协作知识，除了少数情况外，象限1、2和4的研究工作与基于注意力的基线相比，可以获得令人满意的推荐性能。在这些研究中，虽然我们可以观察到适应的LLM的大小随着时间轴逐渐增加，但它们之间的细粒度交叉比较(即统一的基准)仍然是空白的。因此，很难直接得出更大的LLM模型规模一定能给推荐系统带来更好的结果的结论。这就引出了一个悬而未决的问题:对于推荐系统来说，更大的语言模型总是更好吗?或者，将小规模语言模型与协作知识注入结合使用是否足够好?我们对这个问题的看法有两方面:•与小规模语言模型相比，大型语言模型在某些需要推理能力的特定任务中仍然是不可替代的。例如，文本特征增强，类人用户交互和对话，以及推荐管道控制。在这些场景中，通常需要涉及LLM而不是小规模LM，以确保任务完成和推荐性能。

•当在RS中扮演同样的角色时(例如，特征编码器)，一般来说，LLM可以比小规模LM获得更好的性能是一个常识。然而，为了在性能增强和计算成本之间取得平衡，小规模LM将是一种更经济的选择。或者说，LLM带来的额外计算成本是否值得性能收益还没有得到很好的验证，特别是当小规模LM作为轻量级替代品时。

**3.LLM擅长对硬样本进行重新排序**。虽然LLM通常在零/少次学习方面表现较差，因为很少涉及领域内的协作知识，但研究人员[62,133]发现，大型语言模型(如ChatGPT)更有可能成为硬样本的良好重新排序器。他们引入了过滤器-再排序范式，利用传统推荐系统的预排序功能(例如，工业应用中的匹配或预排序阶段)来预过滤那些容易的负面项目，从而生成一组具有较难样本的候选项目，供LLM重新排序。这样一来，LLM可以推广类似chatgpt的api。这一发现对工业应用具有指导意义，在工业应用中，我们可以要求LLM只处理硬样本，而将其他样本留给轻量级模型以节省计算成本。

**6.结论与未来展望**

总之，大型语言模型由于其广泛的开放世界知识、逻辑和常识推理能力以及对人类文化和社会的理解而表现出令人印象深刻的类人能力[50,207,262]。因此，大型语言模型的出现为llm增强的推荐系统开辟了一个有前途的研究方向。本调查从工业推荐系统的整个管道的角度对法学硕士增强的推荐进行了系统的观察。本文从何处以及如何将LLM适应RS两方面全面总结了大型语言模型适应推荐系统的最新研究进展。

•对于“WHERE”问题，我们分析了LLM在推荐管道的不同阶段可以扮演的角色，即特征工程、特征编码器、评分/排名功能、用户交互和管道控制器。

•对于“HOW”问题，我们分析了训练和推理策略，得出了两个正交的分类标准，即是否在训练期间调整LLM，以及是否涉及CRM进行推理。

还为每个分类法透视图提供了详细的讨论和有见地的开发路径。对于未来的展望，除了我们在第5节已经强调的三个方面(即效率、有效性和道德)，我们想进一步表达我们对大型语言模型和推荐系统相结合的未来发展的希望:•迫切需要一个统一的公共基准来提供合理且令人信服的评估协议，因为

(1)现有作品之间的细粒度交叉比较仍然空缺，

(2)结合LLM的推荐模型的实验结果重现成本相当高且困难。

虽然存在一些LLM增强的RS的基准，但它们通常集中在LLM增强的RS的某个方面，而不借助CRM。因此,一个统一比较LLM对不同推荐流水线阶段(如特征工程、特征编码器)的适应情况仍有待探索。

•为推荐领域定制的大型基础模型，可以接管整个推荐管道的控制。目前，管道控制器阶段涉及LLM的研究工作，一般采用ChatGPT、GPT4等冻结型通用大型基础模型来连接不同阶段。通过构建领域内的指令数据，甚至定制协作知识的模型结构，我们有望获得专门为推荐领域设计的大型基础模型，从而使推荐系统的自动化程度达到一个新的水平。