**通过高级提示词工程优化LLM在电影推荐重排任务中的性能：实现NDCG@10 0.7-0.8目标的策略研究**

**I. 引言：利用LLM与高级提示词工程革新电影推荐重排**

**A. 推荐系统与LLM的演进格局**

传统推荐系统正经历深刻变革，大型语言模型（LLM）的崛起为其注入了新的活力，尤其是在电影推荐重排这类复杂任务中展现出巨大潜力 1。LLM凭借其强大的自然语言理解能力，能够更深入地解析用户评论、电影剧情梗概等文本信息，捕捉用户细微、复杂的偏好 1。重排（Reranking）作为推荐流程中的关键环节，负责对初步候选列表进行优化，以提升个性化程度和相关性，从而改善用户体验 2。文献 3 强调了重排在精炼推荐结果以及平衡准确性、多样性和公平性等多重目标方面的重要性。同时，2 指出LLM通过上下文理解能力捕捉复杂用户兴趣，为重排任务带来了新的可能性。

用户对个性化推荐的期望日益增高，他们不再满足于简单的类型匹配，而是追求能够理解其深层需求（如情感共鸣、主题关联、新颖性探索）的推荐。传统协同过滤或仅基于内容的推荐方法在处理用户评论、电影剧本等非结构化文本中蕴含的丰富语义时常显不足。LLM恰好弥补了这一缺陷，能够从海量文本中提炼出更精准的用户画像和物品特征。重排阶段通常处理的候选项目数量相对较少，这为LLM应用其深度理解能力提供了理想的场景。因此，对更细致入微的个性化需求的追求，正成为推动LLM在推荐重排领域广泛应用的核心驱动力。

**B. 提示词工程在LLM重排性能中的核心地位**

提示词工程（Prompt Engineering）是指在不更新模型权重的前提下，通过设计和优化输入提示（Prompt）来引导LLM行为，以期获得理想输出的方法 5。这是一门经验性科学，需要大量的实验和探索 5。对于电影推荐重排这类需要理解用户微妙偏好、物品多维度特征以及权衡潜在冲突目标（例如，在确保准确性的同时提升多样性）的复杂任务而言，标准或基础的提示词可能难以胜任。文献 7 指出，对于复杂的推理任务，仅靠基本的少样本提示（Few-shot Prompting）可能不足以获得可靠结果，这促使了思维链（Chain-of-Thought, CoT）等更高级技术的出现。

LLM在重排任务中的成功，很大程度上取决于提示词工程的质量。这意味着推荐系统开发者的技能组合正在发生演变，除了传统的机器学习模型训练能力外，与LLM进行高效“沟通”的能力也变得至关重要。LLM功能强大，但并非能够“心灵感应”。它们在特定任务（如重排）上的表现，直接取决于任务目标和约束条件是否被清晰、准确地传达。拙劣的提示词可能导致次优、泛化甚至不相关的重排结果。高级提示词技术，本质上是阐述复杂重排逻辑（例如，平衡准确性与多样性、推断用户未明确表达的偏好）的精密方法。因此，提示词工程正成为一项核心竞争力。此外，人工构建提示词不仅耗时，且往往难以达到最优效果，这凸显了对系统化方法和自动化优化策略的需求 2。

**C. 报告目标与范围**

本报告的核心目标是深入研究高级提示词工程技术，特别是思维链（CoT）、表格链（Chain-of-Table）等方法，如何能够有效优化LLM在电影推荐重排任务中的性能。

报告将明确探讨旨在将归一化折扣累计收益（NDCG@10）提升至0.7-0.8区间的策略。这一性能指标代表了推荐系统领域的高水平表现。例如，文献 9 提及通过将自适应检索与LLM重排器结合，nDCG@10最多可提高13.23%；文献 8 展示的AGP框架在特定条件下达到了0.705的NDCG@10；而 10 指出RankGPT结合GPT-4o在DL19数据集上实现了0.7506的nDCG@10。这些数据为本报告设定的目标范围提供了实际背景。

追求0.7-0.8这样高的NDCG分数，标志着推荐系统领域日趋成熟，微小的性能提升也需要日益复杂和精密的专门技术。这一目标推动着研究超越简单的相关性匹配，深入到更深层次的用户偏好理解和列表优化层面。最初，LLM在推荐领域的应用可能通过基础提示词就展现出了一定潜力。然而，要达到与专用模型相媲美甚至超越的专家级性能（如NDCG 0.7-0.8），则需要充分挖掘LLM的潜力。这必然要求采用能够引导LLM进行复杂推理、处理多重目标并实现高效个性化的先进提示词技术。可以说，这一性能目标本身就成为了提示词工程创新的催化剂。

本报告的范围将聚焦于提示词工程方法论，探讨其理论基础、在电影推荐重排场景下的实际应用，以及支持这些技术的相关框架。

**II. 基于LLM的重排方法基础**

**A. LLM作为重排器：处理流程与评估**

LLM在执行重排任务时，通常采用列表式（Listwise）方法，即提示LLM对给定的候选项目列表进行重新排序 9。这要求LLM在特定上下文中理解整个项目集合。输入给LLM的信息通常包括用户的交互历史（如观看过的电影、评分）、候选电影列表（通常由一个初步的检索模型生成），以及可能存在的用户明确查询或偏好 1。LLM的输出则是一个重新排序的电影列表，有时还会附带解释或置信度分数 11。

LLM在重排任务中的有效性，与初始候选列表的质量紧密相关。如果高度相关的项目未能出现在输入列表中，那么无论重排算法多么精妙，也无法将它们呈现给用户。文献 9 强调了这种“有界召回问题”（bounded recall problem）。LLM重排器是在一个预先筛选的列表上进行操作。如果这个列表质量不高（例如，遗漏了用户可能非常喜欢的关键影片），那么LLM的任务就从优化变成了“亡羊补牢”。这意味着端到端的系统视角至关重要；孤立地优化重排器，如果检索阶段薄弱，可能会导致收益递减。因此，针对重排的高级提示词策略，理想情况下应辅以改进候选池的策略。

**B. 捕捉用户偏好与物品特征**

LLM能够从用户的历史交互中总结出其偏好，提取如喜欢的类型、导演、演员、情绪氛围甚至角色类型等信息 1。例如，UR4Rec框架利用LLM生成用户偏好的文本描述 1。同时，LLM也能生成关于物品的丰富描述或知识，超越了基本的元数据 1。

LLM处理和综合非结构化文本数据（如用户历史、物品描述）并将其转化为结构化偏好（如 1 所述）的能力，是其关键的差异化优势。然而，这一过程本身需要通过精心的提示词进行引导，以避免提取虚假或过于笼统的偏好。传统推荐系统依赖结构化元数据或明确的用户反馈。LLM则能从原始文本中推断偏好，例如“用户喜欢对话诙谐、女性主角复杂的电影”。这种能力非常强大。但是，如果缺乏优质的提示词，LLM可能会关注到一些肤浅的方面，或者生成过于宽泛的总结，比如当用户的偏好是特定子类型时，仅概括为“用户喜欢动作片”。因此，用于*用户画像生成*的提示词（如AGP框架中的应用 2）与重排提示词本身同等重要。

然而，用户历史和物品描述等文本数据往往是“嘈杂”的，可能包含不相关信息。有效的提示词对于引导LLM关注显著特征至关重要 2。

**C. 基准比较与高级方法的必要性**

常见的基准方法包括随机推荐（Random）、最流行推荐（MostPop）和基于用户浏览历史的主题流行推荐（TopicPop）12。虽然LLM通常优于这些基准，但目标是显著超越简单的LLM应用。

直接应用LLM进行推荐存在局限性。若无精密的提示词设计或针对性的微调，LLM可能生成冗余或不相关的信息，或者难以准确排序 1。文献 1 指出，LLM并非天生就为推荐和排序任务数据集训练。

LLM生成的提示词在某些情况下能够填补性能差距。文献 12 表明，由LLM生成的提示词往往比人工精心设计或静态的提示词表现更好，这为性能提升指明了一条路径。

尽管列表式重排是LLM常用的方法 9，但LLM的上下文长度限制在处理较长的候选列表时可能成为瓶颈 4。重排任务常常涉及评估数十个项目。如果每次重排决策都将所有项目细节和完整的用户历史输入LLM，可能会超出上下文长度限制，或者计算成本过高。这推动了对更强大的LLM的需求，或者更智能的提示词策略，这些策略能够有效地在上下文窗口内总结或优先处理信息。可能的解决方案包括：(1) 两阶段LLM重排（使用一个轻量级LLM进行初步筛选，然后用一个重量级LLM对较短的列表进行最终重排）；(2) 设计提示词，指示LLM关注项目中最具区分性的特征；(3) 在将用户历史或项目集输入最终重排提示词之前，采用有效总结这些信息的技术。

**III. 面向最优重排的高级提示词工程技术**

**A. 思维链（CoT）提示：实现多因素推理**

思维链（Chain-of-Thought, CoT）提示通过引导LLM在得出最终答案前，生成一系列中间推理步骤 5。这种方法对于需要逻辑推演或综合考量多个因素的复杂任务尤其有效 5。

在电影推荐重排中的应用主要体现在以下方面：

1. **整合多重考量维度**：LLM4Rerank框架利用CoT来综合考虑准确性、多样性、公平性等多个方面，将这些维度抽象为图中的节点。LLM根据当前的重排历史和一个称为“目标（Goal）”的输入来决定接下来要考虑的节点（即考量维度）3。这种机制使得对候选列表的评估更为全面。
2. **分步执行重排逻辑**：一个CoT提示可以指示LLM按以下步骤操作：
   * “首先，评估每部电影与用户明确表述的偏好（例如‘具有宏大世界观的科幻片’）的相关性。”
   * “其次，考虑用户观看历史所体现的隐式偏好（例如‘用户经常观看导演X的作品，即使该类型并非其声称的最爱’）。”
   * “再次，评估排名靠前电影的多样性。是否存在过多来自同一子类型或系列的影片？”
   * “最后，综合以上考量，生成最终的重排列表。”

为了阐释CoT提示的结构，可以参考少样本CoT提示的示例 5，并针对电影推荐重排场景进行定制，展示如何构建推理步骤。例如，演示如何平衡用户对“动作片”的偏好与“高分口碑电影”及“近期上映”等因素。

CoT的优势在于其推理过程更为透明，能更好地处理复杂权衡，并可能带来更强的泛化能力。文献 5 指出，CoT对于复杂的推理任务和参数规模较大的模型效果尤为显著。

**B. 表格链（Chain-of-Table）：结构化推荐推理过程**

表格链（Chain-of-Table）是对CoT的扩展，它在推理链中使用动态演变的表格作为中间思考过程的代理 13。LLM通过迭代生成操作（如添加列、选择行、分组）来转换初始表格，使得推理过程更加结构化，中间结果也更加明确 13。

将表格链应用于电影推荐重排，可以设想以下步骤：

1. **初始表格构建**：
   * *用户画像表*：行代表用户观看过的电影，列包含电影标题、类型、导演、演员、用户评分、观看日期、衍生用户关键词（例如“烧脑剧情”、“大女主”）。
   * *候选电影表*：行代表待推荐的候选电影，列包含电影标题、类型、剧情梗概关键词、平均评分、上映年份、导演、演员。
2. **用于重排逻辑的表格操作**：
   * FILTER\_PROFILE\_BY\_RECENT\_ACTIVITY：根据观看日期从用户画像表中筛选近期观看记录。
   * EXTRACT\_KEY\_PREFERENCES：从（筛选后的）用户画像中识别主要的类型、导演、演员、衍生用户关键词。将这些信息存储为一个新的小型“偏好”表，或作为列添加到工作表中。
   * SCORE\_CANDIDATES\_RELEVANCE：对候选电影表中的每部电影，根据其类型、导演、演员、剧情梗概关键词与提取出的核心偏好信息的匹配度计算相关性得分，并添加为相关性得分列。
   * GROUP\_BY\_GENRE\_AND\_DIVERSIFY：按类型对候选电影进行分组。如果基于相关性得分，某个类型的电影在Top N中占比过高，则应用降权因子或从每个类型中选择Top M。这可能涉及添加多样性得分列或排序调整标记列。
   * FINAL\_SORT：基于综合得分（例如相关性得分、多样性得分、平均评分的加权和）对候选电影进行最终排序。

LLM的角色是（通常通过少样本学习）根据整体重排目标（例如“最大化Top 10电影的相关性，同时确保类型多样性”）生成这些操作及其参数的序列。LLM本身不直接执行类似SQL的操作，而是生成这些操作的描述，然后由程序化方式执行 13。

表格链的优势在于为复杂推理提供了增强的结构，通过明确的中间结果提升了可解释性和调试便利性，并可能更准确地处理涉及结构化数据的多步骤决策过程 13。表格链中的“操作” 13 虽然是为通用表格操作定义的，但在电影推荐重排场景下，这些操作需要进行领域适配。其真正的威力在于定义一套与推荐场景中常见推理步骤相符的操作集（例如，“查找相似项目”、“检查系列完整性”、“平衡探索与利用”）。LLM的任务将是学习如何对这些领域特定的操作进行排序。这使得表格链具有高度的可扩展性，但也要求对操作集进行精心设计。

**C. 少样本提示（Few-Shot Prompting）：实现上下文学习与个性化**

少样本提示通过向LLM提供少量输入输出对示例（“shots”），来演示期望的任务、输出格式、语气和风格 5。这种方法利用了LLM的上下文学习能力 17。

在电影推荐重排中的应用包括：

1. **演示重排逻辑**：提供候选列表、用户画像片段以及期望的重排结果，并简要解释为何某些电影被提升或降低了排名。
   * *示例*：
     + 用户画像：“喜欢克里斯托弗·诺兰的电影，偏爱心理惊悚片，近期观看过《盗梦空间》、《信条》。”
     + 初始候选列表：
     + 期望重排列表：
     + 解释：“优先推荐《记忆碎片》和《星际穿越》，因为符合导演偏好和类型匹配。‘动作片A’因大众吸引力保持较高排名。其他影片优先级降低。”
2. **控制输出风格**：如果LLM还需要为其重排结果生成解释，少样本示例可以展示解释的期望详细程度、语气（例如，分析型、热情型）和结构 17。
3. **处理特定情况**：使用示例来展示如何处理边缘情况，例如用户偏好稀疏或存在矛盾时。

选择示例时需注意 5：质量优于数量，高质量、有代表性的示例至关重要；语义相似性，选择与测试用例语义相似的示例 5；警惕偏见，注意示例呈现中的多数标签偏见和近因偏见 17，示例中标签的分布也很重要 7。

少样本提示的优势在于能够显著提升LLM在复杂任务（零样本学习可能不足以应对）上的性能 7，并在无需微调的情况下帮助LLM的输出与特定需求对齐。

**D. 其他潜在高效的提示策略**

1. **指令提示（Instruction Prompting）与角色分配（Persona Assignment）**：
   * 清晰地向LLM下达关于任务、期望输出和约束条件的指令 5。
   * 为LLM分配一个角色，例如：“你是一位资深的影评人和推荐专家。你的任务是为具有以下特征的用户重新排列这份电影列表……” 19。这有助于LLM采用更一致、更合适的风格和专注点。文献 20 建议提示词应像剧本一样设定场景，提供角色、动机和背景。
2. **自洽性采样（Self-Consistency Sampling）**：
   * 在温度（temperature）大于0的设置下生成多个输出或推理链，然后从中选择最佳的一个（例如，通过多数投票或验证指标）5。对于重排任务，这可能涉及生成多个重排列表，并使用多样性或置信度指标来选择最终列表。
3. **自动化提示优化（APE / AGP类方法）**：
   * 利用LLM来生成和优化提示词 5。RecPrompt框架使用基于LLM的优化器 12。
   * AGP（Auto-Guided Prompt Refinement）框架专注于优化用户画像生成提示词，而非直接的重排提示词，它利用基于位置的反馈进行优化 2。这种基于可操作反馈的迭代式优化是一个强大的概念。像AGP 2 和RecPrompt的优化器 12 这样的技术引入了一种“元提示（meta-prompt）”或“观察指令（observation instruction）”，用于指导LLM*优化*另一个提示词。这是提示词工程中一个更高层次的抽象。我们不仅提示LLM执行任务（重排），还提示另一个LLM（或同一LLM在不同步骤中）*改进*用于重排任务的提示词。这种自我校正或自动化改进循环 2 对于扩展提示词工程、超越手动试错至关重要，尤其是在追求高NDCG目标时。
4. **检索增强生成（RAG）原则获取上下文**：
   * 虽然RAG本身不是一种提示*类型*，但确保提示词能够访问最丰富、最相关的信息（例如，详细的电影元数据、用户评论、相似用户画像）至关重要。重排可以被视为一种专门的RAG任务。文献 21 讨论了RAG的改进策略。

高级提示词技术并非相互排斥，它们可以有效地结合起来发挥更强大的作用。例如，表格链方法可以使用少样本示例来教导LLM*如何*选择和参数化表格操作。LLM4Rerank的CoT中的“目标”输入 3 本身也可以通过类似AGP的机制进行优化。CoT提供了推理结构，表格链在推理过程中提供了结构化数据操作框架，少样本学习则为执行CoT或表格链中的步骤提供了具体示例，而角色分配设定了整体的上下文。这表明对于复杂任务，采用分层或组合式的提示词设计方法可能更为有效。

CoT和表格链使LLM的推理过程更加明确。这不仅有利于提升性能，还有助于增强推荐的可调试性、可解释性和可信度——这些品质正日益受到用户和系统开发者的重视。如果LLM生成一个没有任何解释的重排列表，它就是一个黑箱。但如果它提供了CoT过程或展示了表格链的转换步骤，开发者就能理解*为何*会产生某个特定的列表，识别逻辑中的错误，而用户如果能看到基本原理，也可能更信任推荐结果 11。这种明确性是其相较于不透明模型的关键优势。

**表1: 高级提示词工程技术对比分析**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **技术名称** | **核心机制** | **在电影推荐重排中的应用价值** | **NDCG@10提升潜力** | **关键文献参考** |
| 思维链 (CoT) | 引导LLM生成中间推理步骤，再得出最终答案。 | 平衡准确性、多样性、新颖性等多重目标；处理复杂的用户偏好和物品属性。 | 高 | 3 |
| 表格链 (Chain-of-Table) | 使用动态演变的表格作为中间思维的代理，通过表格操作进行结构化推理。 | 对用户历史、候选物品进行结构化分析、筛选、评分和排序，使复杂逻辑更清晰。 | 高 | 13 |
| 少样本提示 (Few-Shot) | 提供少量输入输出示例，使LLM通过上下文学习理解任务。 | 演示期望的重排逻辑、输出格式（如带解释的列表）、特定场景处理（如冷启动用户）。 | 中至高 | 17 |
| 角色分配 (Persona) | 为LLM设定一个特定角色（如“电影专家”），引导其输出风格和专注点。 | 使LLM输出更符合推荐场景的专业性和用户期望的语气。 | 中 | 19 |
| 自洽性采样 | 多次生成输出并选取最优，增强结果的鲁棒性。 | 生成多个候选重排列表，通过评估指标（如综合考虑相关性和多样性）选择最终推荐。 | 中 | 5 |
| 自动化提示优化 (APE/AGP) | 利用LLM或其他机制自动生成、评估和迭代优化提示词。 | 持续改进用户画像生成提示或重排提示，以适应数据变化和提升长期性能，减轻人工调优负担。 | 高 | 8 |

**IV. 实现目标NDCG@10的框架与方法论**

**A. AGP (Auto-Guided Prompt Refinement) 个性化重排框架** 2

AGP框架的核心思想是优化用户画像生成提示词，而非直接修改重排提示词。通过生成更优质的用户画像，LLM能够更好地捕捉个性化兴趣，从而实现更有效和可解释的重排 2。

其关键机制包括：

1. **基于位置的反馈（Position-Based Feedback）**：AGP不单纯依赖NDCG等聚合指标，而是利用项目级别的排序错位（例如，“项目X当前排名第5，理想排名应为第2”）来生成可操作的反馈，用于优化提示词 2。这种反馈更具可解释性，并提供了直接的改进信号。这种具体反馈可能比单独使用NDCG等聚合指标更能加速提示词优化过程，并以更少的数据实现更好的泛化。告诉系统“你的NDCG@10是0.6”远不如“这部特定的电影本应排名更高，原因在于（基于真实情况的隐含信息）”来得有帮助。基于位置的反馈提供了更细致的错误信号，使提示词优化器能够进行更有针对性的调整。这类似于为何逐参数梯度比单一损失值对训练神经网络更有效。AGP的数据效率 8 也支持了这一点。
2. **批量训练与汇总反馈（Batched Training and Summarized Feedback）**：AGP跨多个用户聚合反馈，以减轻对个体用户特质的过拟合，确保提示词更新具有广泛适用性 2。
3. **迭代优化（Iterative Refinement）**：用户画像提示词在多个迭代过程中以类似梯度的方式进行优化 2。

性能方面，AGP仅用100个训练用户就在基准模型上实现了5.61%至20.68%的NDCG@10提升，最高达到0.705 2。该框架在图推荐模型上表现尤为出色 2。将训练用户从100增加到700时，NDCG@10仅从0.696略微提升至0.705，显示出其数据高效性 8。

**B. LLM4Rerank: 基于CoT的多维度重排框架** 3

LLM4Rerank的核心思想是利用零样本LLM和CoT在重排过程中整合准确性、多样性、公平性等多个方面 3。

其关键机制包括：

1. **维度节点（Aspect Nodes）**：不同的考量维度被表示为全连接图中的节点 3。
2. **CoT导航（CoT for Navigation）**：LLM使用CoT来决定考虑维度的顺序，这一过程受到当前重排历史和一个“目标（Goal）”输入的影响 3。这个“目标”输入允许对重排的侧重点进行个性化定制。
3. **可扩展性（Scalability）**：新的考量维度可以作为新节点加入框架 4。

性能方面，LLM4Rerank在三个广泛使用的公开数据集上，在多个维度均超越了现有的SOTA重排模型 3。文献 10 提及LLM4Ranking（一个相关框架，可能是LLM4Rerank的演进版本或笔误）使用GPT-4o实现了0.7506的nDCG@10。LLM4Rerank利用强大的零样本LLM 3，而AGP则专注于优化提示词，这可能适用于任何LLM，甚至能通过高度优化的用户画像使小型模型表现良好。这提供了一个战略选择：是依赖大型模型的原始能力，还是投资于精密的提示词优化框架。AGP达到的0.705的NDCG@10 8 具有竞争力，表明后一种方法同样具有价值。

**C. UR4Rec: 基于用户偏好检索的增强型重排框架** 1

UR4Rec的核心思想是通过从LLM生成的用户/物品知识中检索必要信息，并将LLM与传统推荐器对齐，从而增强重排效果 1。

其关键机制包括：

1. **LLM生成的知识（LLM-Generated Knowledge）**：LLM生成用户偏好（如类型、导演、情绪）和物品知识的文本摘要 1。
2. **用户偏好检索器（User Preference Retriever）**：这是一个基于Transformer的组件，它以候选物品作为查询，从生成的知识中筛选并检索相关的用户偏好 1。这在LLM输出和推荐模型表示之间架起了桥梁。
3. **预训练目标（Pre-training Objectives）**：采用对比学习和偏好-物品匹配任务，以将LLM知识与推荐模型的嵌入对齐 1。

虽然UR4Rec并非直接针对最终重排步骤的“提示词工程”技术，但它强调了为重排器提供高质量、经LLM处理的输入（用户偏好）的重要性。用于生成这些初始用户/物品知识的提示词至关重要。

**D. RecPrompt: 提示词工程框架** 12

RecPrompt是一个包含提示词优化器、推荐器和监控器的框架，旨在增强新闻推荐（可适用于电影推荐）的效果。

其关键机制包括：

1. **提示词优化器（Prompt Optimizer）**：通过整合系统消息、当前提示词、推荐样本和一个“观察指令”（元提示），来优化初始提示词模板，以指导LLM调整提示词 12。
2. **反馈循环（Feedback Loop）**：创建反馈循环，其中生成的推荐和提示词性能（例如MRR）会反过来用于进一步优化提示词 12。
3. **LLM生成与人工构建对比**：LLM生成的提示词往往比人工构建的提示词表现更好 12。

性能方面，LLM生成的提示词相比初始静态提示词平均带来了约4-6%的性能提升 12。

多个框架（如AGP、RecPrompt）都强调利用LLM和反馈循环进行*自动化提示词优化*。这是一个强劲的新兴趋势，表明尽管手动设计提示词是基础，但为了达到SOTA性能，它很可能被自动化方法增强或取代。手动找到“完美”的提示词极其困难，因为搜索空间巨大，且提示词元素之间的相互作用复杂。AGP 2 和RecPrompt 12 等框架承认了这一点，它们利用LLM本身，在性能反馈的引导下，探索和优化提示词。这本质上是将机器学习应用于提示词设计问题。

同时，AGP（优化用户画像提示）和UR4Rec（检索/优化LLM生成的用户偏好）等框架凸显了输入到最终重排LLM的*信息质量*与重排提示词本身同样关键。“输入决定输出”的原则依然适用。重排LLM基于其获得的关于用户和物品的信息进行操作。如果用户画像未能准确概括其真实偏好（AGP的关注点 2），或者物品特征未能得到良好表征，那么无论其提示词多么精妙，重排器都无法发挥最佳性能。UR4Rec的检索器 1 旨在提供清晰、相关的偏好信号。这暗示多阶段提示策略可能是最优选择。

**表2: LLM重排框架及其性能洞察**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **框架名称** | **核心提示策略/创新点** | **性能关键特征** | **已报告的NDCG@10 (及上下文)** | **相关文献参考** |
| AGP | 优化的用户画像生成提示 + 基于位置的反馈 | 迭代优化，批量训练，精确的排序校正信号 | 高达 0.705 (Movies & TV, LightGCN, 100训练用户) | 2 |
| LLM4Rerank | 基于CoT的多维度重排 (准确性、多样性、公平性等) | 动态维度考量，通过“Goal”输入实现个性化，可扩展性强 | 超越SOTA基准 (具体数值未统一给出，但效果显著)；相关框架LLM4Ranking使用GPT-4o达0.7506 | 3 |
| UR4Rec | LLM生成用户/物品知识 + Transformer检索器对齐偏好 | 增强传统推荐器的输入，桥接LLM与推荐模型语义鸿沟 | (未直接报告NDCG@10，但旨在提升下游重排性能) | 1 |
| RecPrompt | LLM驱动的提示词优化器 + 反馈循环 | 自动化提示词迭代，整合多种输入进行提示词调整 | LLM生成提示比静态提示平均提升4-6% (MRR等指标) | 12 |

**V. 实现NDCG@10 0.7-0.8的实践路线图与考量**

**A. LLM选择与配置**

选择LLM时需考虑模型规模、推理能力（CoT受益于大型模型 5）、上下文窗口长度、推理成本以及微调选项。开源模型与基于API的模型各有优劣，LLM4Ranking框架同时支持这两种类型 10。DeepSeek-V3是一款强大的开源MoE（Mixture-of-Experts）模型 23，如果需要精细控制或看重其特定的架构特性（如其MoE结构带来的效率/容量平衡），则可能是一个合适的选择。其架构 24 旨在通过增加稀疏性和模型规模，在相似的推理成本下实现更高质量。这表明模型的设计正朝着更好地适应复杂提示结构或提升“可提示性”的方向发展。最佳的LLM是在预算和延迟约束下，能够最好地响应所选高级提示词技术的模型。此外，温度（temperature）、最大令牌数（max\_tokens）等参数设置也需要通过实验来确定 25。

**B. 面向提示词的数据准备与特征工程**

用户历史数据如何在提示词中简洁而全面地呈现（例如，电影标题列表、类型、UR4Rec中派生的关键词 1，或AGP中生成的画像 2）至关重要。对于候选物品，提示词必须包含足够的信息（标题、类型、导演、演员、剧情梗概、用户评分、关键词），以便LLM做出明智决策。在构建少样本示例时，应选择多样化且具代表性的例子，覆盖各种场景和用户偏好 5，甚至可以利用LLM辅助生成示例 17。

在提示词中如何*表示*用户历史和物品信息，是提示词工程中一个关键且常被忽视的方面。这不仅仅涉及提示词的指令部分，还包括数据的“有效载荷”。LLM只能基于其上下文中提供的数据进行推理。如果用户历史仅仅是一串电影ID，其效用远不如电影标题列表，甚至不如标题+类型+用户生成的标签/关键词。物品信息同理。AGP中的“用户画像生成提示” 2 和UR4Rec中LLM生成的用户偏好 1，其核心都是为下游重排LLM创建一个丰富、凝练的用户/物品数据表示，使其能有效利用。这可以视为一种针对LLM提示词的“特征工程”。

**C. 提示词设计与迭代策略**

首先应从清晰的基于指令的提示和明确的角色设定入手 5。然后逐步整合高级技术：

1. 引入少样本示例以实现基本个性化。
2. 开发CoT提示以进行多维度推理（例如，相关性、多样性、新近度）。
3. 如果结构化数据操作是期望排序逻辑的核心，则探索表格链。

提示词设计是一个经验性的迭代过程 5。应使用验证集测试不同提示词的变体，并持续追踪NDCG@10及其他相关指标。如果条件允许，可以实施或调整类似AGP 2 或RecPrompt 12 的框架，基于反馈持续优化提示词。

即使采用自动化提示词优化（APE、AGP），在初始阶段，由领域专家（理解何为优质电影推荐）指导的手动提示词设计和迭代，可能仍然不可或缺。RecPrompt中的“观察指令” 12 或LLM4Rerank中的“目标” 3 通常需要人类的洞察力。自动化系统可以基于某个指标优化提示词，但定义提示词的初始搜索空间、提示词结构以及高层目标，往往需要人类智慧。例如，决定在LLM4Rerank的CoT提示中包含哪些维度（准确性、多样性、新颖性、惊喜度），是一个需要人工判断的决策。人类设定战略方向，自动化系统则负责微调提示词的战术执行。文献 12 提及，结合人类与LLM进行提示词优化能取得最佳效果。

**D. 评估方法**

主要评估指标为NDCG@10。次要指标可包括：

* 召回率（Recall）9 – 确保相关项目获得高排名。
* 多样性（Diversity）（例如，列表内相似度、类型分布）。
* 公平性（Fairness）（如适用，例如不同类型电影的曝光机会）。
* 可解释性（Explainability）（定性评估，或者如果生成了解释，则评估其质量）。
* 计算成本/延迟。

基准比较方面，应与强大的基准模型进行对比，包括非LLM方法和简单的LLM提示方法 8。文献 21 强调对变更进行基准测试，并争取至少5-10%的NDCG提升以证明实施成本的合理性。

**E. 集成用户反馈回路**

除了AGP框架中基于位置的反馈外，还应考虑整合真实用户的反馈（显式反馈：评分、点赞/点踩；隐式反馈：点击、观看时长），以便进一步优化提示词，甚至随时间推移在个体层面实现个性化 21。

要达到NDCG@10目标区间的上限（0.75-0.8），可能需要采用计算密集型技术（例如，超大型LLM、包含许多步骤的复杂CoT、多次采样的自洽性验证）。文献 10 提到RankGPT与GPT-4o（一个非常大的模型）实现了0.75的NDCG。更复杂的推理（CoT、表格链）本身就需要LLM进行更多处理。自洽性验证会成倍增加推理成本。尽管用户的首要目标是NDCG，但实际部署时会面临预算和延迟的限制。因此，实践路线图必须承认这一点，并提出优化这种权衡的策略，例如通过模型量化（10 提及bitsandbytes、GPTQ）、知识蒸馏或提示词压缩。

**VI. 结论：综合策略以实现高性能电影推荐重排**

**A. 关键提示词工程技术及其协同作用回顾**

本报告探讨了多种高级提示词工程技术，包括思维链（CoT）、表格链（Chain-of-Table）、少样本提示、角色分配以及自动化提示词优化。这些技术并非孤立存在，而是可以相互结合，协同增效，共同致力于提升LLM在电影推荐重排任务中的性能。例如，可以使用少样本示例来引导CoT的推理步骤，或者为执行表格链操作的LLM赋予一个专家角色。这种组合策略能够更全面地发挥LLM的潜力。

**B. 通往NDCG@10 0.7-0.8之路**

实现NDCG@10 0.7-0.8这一具有挑战性的目标，需要一个整体性的综合方案。这不仅仅依赖于某个“神奇的提示词”，更是一个系统工程。它要求选用能力强大的LLM，配合精心设计乃至自动优化的提示词（以引导CoT或表格链等复杂推理），确保输入高质量的数据（包括丰富的用户画像和物品详情），并建立鲁棒的评估体系。AGP和LLM4Rerank等框架（或类似的定制化实现）为此提供了结构化的解决路径。文献中报告的NDCG值 8 表明，尽管目标宏大，但借助前沿方法是可能实现的。整个系统的设计，从数据表征、提示策略、LLM能力到反馈机制，必须协同工作。任何一个环节的薄弱（例如，糟糕的用户画像生成）都可能削弱其他环节的优势（例如，一个出色的CoT重排提示）。

**C. 未来研究方向与待解决问题**

尽管LLM和高级提示词工程已展现出巨大潜力，但仍有广阔的研究空间：

* **自动化多目标提示词优化**：进一步发展能够同时优化NDCG、多样性、可解释性等多个重排目标的自动化提示词优化技术。
* **领域定制的表格链操作**：为推荐推理设计更复杂、更贴合业务逻辑的表格链操作。
* **LLM架构与提示词的交互**：深入理解不同LLM架构（如DeepSeek-V3等MoE模型 23）与复杂提示词结构之间的相互作用机制。
* **动态提示词自适应**：研究能够根据实时用户上下文或物品特性动态调整提示词的技术。
* **LLM驱动的智能体在交互式推荐中的应用**：探索LLM驱动的智能体在更具对话性和交互性的推荐场景中的潜力 26。
* **提升LLM重排器的可解释性**：利用CoT或表格链生成的推理过程，增强LLM重排结果的可解释性 11。

该领域发展迅速，今天的SOTA技术可能很快被超越。因此，掌握良好提示词工程的原则（清晰性、上下文、结构化推理、迭代优化）比任何单一特定的提示词模板更为持久。用户对0.7-0.8 NDCG的追求只是时间长河中的一个片段。本报告旨在提供基础性的策略，以便在新的LLM和提示词范式出现时能够灵活适应。

最后，虽然NDCG@10是明确的性能目标，但高级提示词技术（如LLM4Rerank中用于提升多样性/公平性的CoT 3）以及其在可解释性方面的潜力 11，共同指向了对推荐质量更广泛的定义。追求高NDCG可以，也或许应该，与改善这些其他以用户为中心的维度相结合。真正的“优化”目标是整体用户体验，而NDCG是衡量这一体验的重要指标之一。