LARF：面向搜索相关性的学习对齐推理框架

Tian Tang[1], * Guo Chen[1,2],* Reed Li[1],† Hao Wang[1] Yu Lu[1],†
Haijin Liang[1] Haibo Shi[1]
[1]元宝团队，腾讯 [2]西南大学

arXiv:2611.32125v1 [cs.IR] 24 Nov 2026

摘要

查询-文档相关性估计是原生 AI 搜索系统的核心组成部分，模型必须依据复杂的多维标准，评估检索到的文档是否满足用户的信息需求。然而，将大语言模型（LLM）直接应用于这一任务仍然颇具挑战，因为传统的监督微调（SFT）对领域专属判断依据提供的监督有限，而标准的轨迹级强化学习（RL）目标则不会显式地约束相邻有序相关性标签之间的间隔。此外，尽管这类推理在训练过程中可能很有用，但思维链（CoT）推理并不适用于对延迟敏感的在线排序。这些挑战使模糊决策边界难以厘清。我们提出了一个三阶段框架 LARF（面向搜索相关性的学习对齐推理框架）：首先，针对策略性选取的样本，利用多智能体引导的思维链依据进行冷启动监督微调；随后，应用判别式GRPO，结合按属性、按列表和规则一致性奖励，并对末端相关性词元施加间隔损失；最后，将教师模型蒸馏为双头学生模型，并保留其分类头以进行非自回归推理。在一个人工标注的网页搜索相关性基准上，该基准包含约 30 万个训练样本和 1.6 万个测试样本，LARF-8B 将六分类准确率提高了 5.78 个百分点，将归一化折损累计增益（NDCG）提高了 5.05 个百分点，同时纯分类服务路径将推理延迟降低了 82.9%。这些结果表明，对结构化推理、终端决策优化和分类头蒸馏进行对齐，能够为工业级搜索相关性建模实现有效的准确率–效率权衡。

1 引言

大语言模型（LLM）的最新进展正推动信息检索（IR）从传统的文档匹配转向更广泛的以大语言模型增强的范式，涵盖生成式检索、基于大语言模型的重排序、检索增强生成以及搜索智能体（Zhu 等

，2025年；Li 等，2025年）。基于这些能力，原生 AI 搜索系统整合检索、推理和答案生成，以支持复杂且开放式的信息需求；代表性产品包括 ChatGPT、搭载 Gemini 的 Google 搜索、豆包和元宝（Ope-nAI，2024年；Google，2025年；腾讯，2026年；字节跳动，2026年）。在此类系统中，搜索相关性仍然是决定搜索质量的核心因素，因为它决定检索到的文档是否能为满足用户潜在的信息需求提供有用证据，进而影响下游答案质量、检索精确率和用户满意度（Mao 等，2016年；Tang 等，2025年）。原生 AI 搜索系统会将检索到的证据综合为直接答案，因此，相关性估计对于排序质量以及生成响应的忠实性和实用性都至关重要。

然而，一个尚未解决的核心问题是，如何构建基于大语言模型的相关性模型，使其能够学习领域特定判断规则、可靠地做出细粒度相关性判断，同时保持足够高的效率，以适用于在线搜索系统。监督微调（SFT）将查询-文档对映射为相关性标签（Yuan 等，2023年；Qiao 等，2019年；Wiseman 和 Rush，2016年）。仅标签监督微调能够学习输入-输出映射，但对于相关性判断所依据的推理依据，提供的监督有限（Xia 等，2024年；Chang 等，2024年；Lu 等，2025b）。传统的强化学习（RL）方法（例如 GRPO）主要优化轨迹级奖励（Zenget al.，2026年），这给搜索相关性建模带来了两个局限。首先，对整个

*共同贡献。 †通讯作者：luyuresearch@foxmail.com，reedsli@tencent.com

推理轨迹进行奖励，可能会助长不必要的冗长推理依据，从而削弱对简洁判别性表示的学习（Renze and Guven，2024年）。其次，由于最终相关性预测由分类标记（例如搜索相关性类别）决定，轨迹级奖励无法为这一关键标记提供明确的细粒度监督。因此，模型仍可能难以区分相邻的相关性等级（Liuetal.，2025年）。尽管思维链结合强化学习可以改善相关性建模，但多步自回归生成会带来显著延迟（Wei etal.，2022年；Chu etal.，2023年；Yeoet al.，2025年；Chen etal.，2026年）。这种延迟往往无法被在线系统接受。

综合来看，这些局限表明，仅仅扩大标签监督规模或优化更长的推理轨迹，并不足以实现搜索相关性建模。首先，推理应作为训练时信号，揭示相关性判断背后的结构化标准，以便从规模较小、经过策略性筛选的语料库中学习判断依据。其次，应针对多个相关性维度显式优化终端序数决策，而不是任由其依赖稀疏的轨迹级奖励。第三，自回归思维链生成会增加服务延迟和成本，因此有必要将其蒸馏为能够支持实时排序的高效学生模型（Chen 等人，2023年；Eigner 和 H¨andler，2024年；Yang 等人，2024年；Xu 等人，2025年）。这些需求亟需一个统一框架，以实现多维相关性学习、显式决策优化和高效服务。

本文提出 LARF（**L**学习对齐**A**推理**R**框架**F**，面向搜索相关性），这是一个将推理导向的学习与高效相关性服务相结合的三阶段框架：**（1）引导式思维链冷启动监督微调** 通过分层多样性采样和困难样本挖掘构建训练子集，使模型能够从信息量最大的边界案例中学习；随后，以多智能体依据扩充人工相关性标签，将判断分解到不同维度；**（2）带边际对齐的判别式 GRPO** 结合点式、列表式和规则一致性奖励，以对齐生成的推理依据与相关性偏好，同时采用间隔目标，显式拉开真实相关性标记与其最强竞争类别之间的距离（Shao 等人，2024年；Rosset 等人，2003）；**（3）双头知识蒸馏** 使用多维分类头和辅助生成式推理头训练共享骨干网络，随后利用分类头进行高效的大规模知识蒸馏（Xu 等人，2025年）。总体而言，LARF 将结构化推理学习、判别式序数决策对齐和高效的仅分类服务整合为一个面向搜索相关性建模的统一训练到部署框架。

我们通过大量实验评估 LARF。在一个人工标注的网页搜索相关性基准上，该基准包含约 30 万个训练样本和 1.6 万个测试样本，LARF-8B 将六分类准确率从 45.20% 提升至 50.98%，并将归一化折损累计增益从 74.51 提升至 79.56。此外，与自回归推理相比，纯分类服务路径将每对样本的推理延迟从 5.68 秒降至 0.97 秒，加速比为 5.86 倍。在 Yuanbao AI Search 的 A/B测试中，LARF 将高满意度率提高了 1.77%，并将低满意度率降低了 3.22%。这些结果表明，通过训练阶段的结构化推理和服务阶段的判别式分类，可以实现切实可行的准确率–效率权衡。

主要贡献总结如下：

- 我们从三个相互关联的挑战出发，分析基于大语言模型的搜索相关性建模：结构化相关性推理、跨层级决策对齐，以及延迟敏感型在线服务。我们将搜索相关性形式化为一种多维有序分类问题，综合考量内容质量、语义相关性、来源权威性和时效性。
- 我们提出了 LARF（**L**earning-**A**ligned **R**easoning**F**ramework for Search Relevance，面向搜索相关性的学习对齐推理框架），这是一种用于搜索相关性建模的三阶段学习对齐推理框架。LARF 包含：（i）引导式思维链冷启动监督微调；（ii）用于终端决策优化的带边际对齐的判别式 GRPO；以及（iii）双头蒸馏，以实现高效的非自回归服务。
- 我们通过大量实验评估 LARF，包括离线评估、组件消融实验和在线 A/B 测试。结果表明，LARF 能够持续提升搜索相关性质量，同时降低服务延迟，并改善用户满意度指标。

2 相关工作

2.1 大语言模型用于搜索相关性建模

搜索相关性建模旨在估计候选文档满足用户信息需求的程度，是检索与排序系统的核心组成部分。早期神经网络方法为查询-文档匹配构建了学习得到的语义表示，突破了词汇匹配的局限，转向判别式相关性估计（Shen 等人，2014；Koroteev，2021）。随着大语言模型的兴起，后续研究探索了基于提示词的相关性预测、任务特定监督适配、数据构建，以及面向搜索相关性建模的推理训练（Zhang 等人，2025c；Xie 等人，2025年）。另一类研究将大语言模型用作相关性评估器，通过将二元、分级和基于偏好的判断与人工标注进行比较，揭示了它们在相关性评估方面的潜力，以及对判断表述方式的敏感性（Faggioli 等人，2023年；Arabzadeh 和 Clarke，2025年）。具体而言，研究表明，在基于大语言模型的排序中，采用细粒度相关性标签而非二元判断，有助于更好地区分部分相关文档（Zhuang 等人，2024a）。近期的工业界研究进一步结合信息性数据选择、结构化理由构建和面向偏好的优化，以提升复杂搜索场景中的相关性判断（Tang 等人，2025年；Lu 等，2025a）。这些进展表明，大语言模型能够超越浅层语义相似度来建模相关性；然而，现有方法对相关性作为多维序数决策问题的处理仍然有限，也未能在单一框架中充分统一结构化相关性学习、对最终相关性判断的直接优化以及高效的在线服务。

2.2 推理增强的大语言模型排序与决策对齐

推理增强的大语言模型排序研究探讨了语言模型如何利用中间推理和对齐目标，改进复杂查询与有歧义的候选文档的相关性判断。基础研究表明，提示大语言模型生成中间推理步骤，可以提升其在需要多步推理任务上的表现（Wei 等，2022年）。与此同时，研究者围绕逐点、成对、列表式和集合式决策形式开发了基于大语言模型的重排序器，展现了在比较和排序候选文档时不同的效果—效率权衡（Qin 等，2024年；Luo 等，2024年；Zhuang 等，2024b）。后续研究考察了显式推理是否有助于文档排序，并引入推理感知训练方案，以提升排序质量和可解释性（Lu 等，2025b；Zhuang 等，2025年；Liu 等，2025年）。近期方法进一步采用强化学习，通过基于结果或基于偏好的奖励信号，优化推理轨迹、排序偏好和规则一致的模型行为（Zeng 等，2026年；Guo 等，2026年；Liu 等，2026年；Zhang 等，2025b）。另一相关方向则将训练时的推理依据监督迁移到高效预测模型中，使推理能够改进学习得到的表示，而无需在推理时进行自回归推理依据生成（Chen 等，2023年；Eigner 和 H¨andler，2024年；Yang 等，2024年；Xu 等，2025年）。然而，现有方法通常优化推理轨迹或相对排序结果，却没有直接建模终端相关性决策的细粒度序数边界，也未能同时解决多维相关性标准、终止决策对齐以及延迟敏感的仅分类服务问题。

3 预备知识

本节将多维查询-文档搜索相关性匹配问题形式化，并介绍用于 LARF 训练的结构化提示语法。

3.1 多维问题形式化

与仅依赖基于嵌入向量或词频的文档相关性相比，整合多维信息对于确定文档相关性至关重要。因此，我们将搜索相关性建模表述为一个多维序数分类问题。对于查询 q，每个候选文档 $d \in \mathcal{D}_q$ 都与文档内容和辅助元数据 m_d（例如来源和发布时间）相关联。我们将完整的模型输入记为x_q，$_d = (q、d、m_d)$，并使用 $\mathcal{G}_q = \mathcal{D}_q$ 表示q的候选文档组。

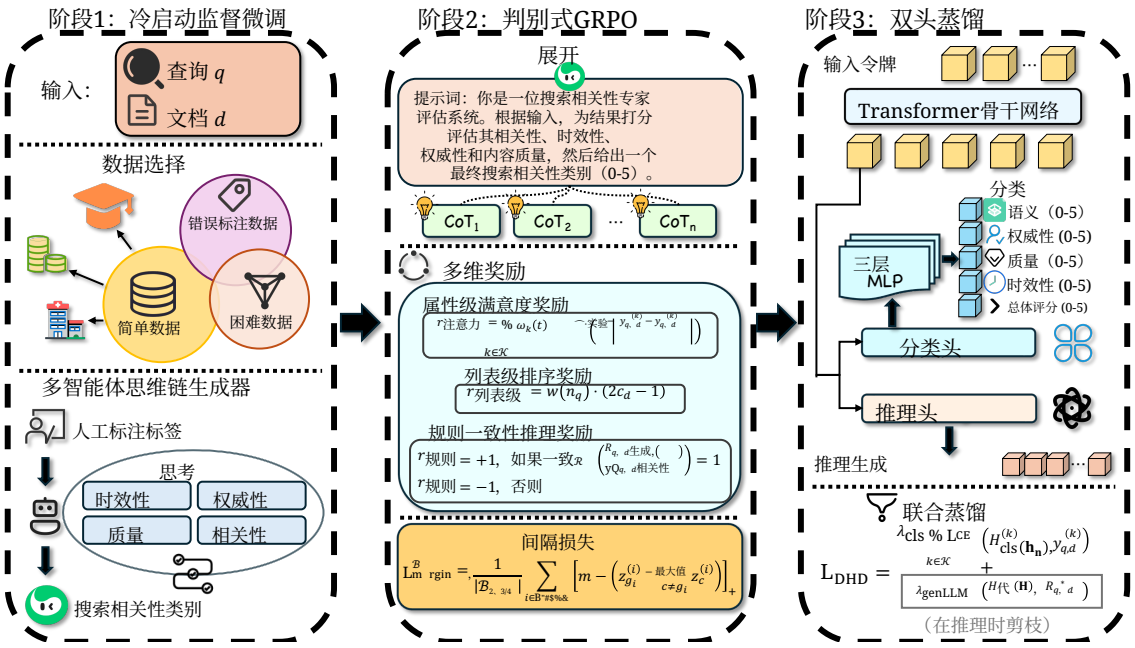

图 1：LARF 的总体流程，包括三个阶段：带引导式思维链的冷启动监督微调、带边际对齐的判别式 GRPO，以及双头知识蒸馏。

每个查询-文档对都标注有一个有序标签向量 $\mathbf{y}_{q,d} = (y_{(k)\,q}、_{d})_{k\in\mathcal{K}}$，其中 $\mathcal{K}$ ={sem、aut、qua、time、rel} 分别表示语义相关性、来源权威性、内容质量、时效性和总体评分。每个维度的取值都来自共同的序数标签空间

$$\mathcal{Y} = \{0,1,2,3,4,5\}, \qquad 0 \prec 1 \prec \cdots \prec 5, \tag{1}$$

其中，终端标签 $y(\text{rel})_{\,q,d} \in \mathcal{Y}$ 的取值范围从 完全不相关 到 完全相关。四个辅助维度为整体判断提供结构化证据；我们不假设存在固定的确定性聚合规则，因为这些维度的相对重要性取决于查询和文档的上下文。内容质量评估文档的完整性与结构；语义相关性衡量意图与主题对齐程度；来源权威性体现来源的可信度；时效性则评估文档是否满足查询的时间意图。对于每个查询-文档对 $d \in \mathcal{D}_q$，模型都会预测一个序数标签向量 $\hat{\mathbf{y}}_{q,d}$。因此，学习包含三个互补目标：准确预测各个相关性维度、保留同一查询下文档的相对排序，以及明确区分终端相关性标签与竞争类别。这些目标促成了属性级奖励、列表级奖励和终端标签间隔损失。

3.2 数据结构与提示词构造

我们将每个输入 $\mathbf{x}_{q,d}$ 序列化为一个基于指令的训练实例，将输入提示词 $\mathrm{P}(\mathbf{x}_{q,d})$ 与结构化目标 $\mathrm{T}_{q,d} = (\mathrm{R}_{q,d},\mathrm{y}_{q,d})$ 配对。提示词包含一个 [系统] 指令，用于指定搜索相关性评估器的角色、查询-文档内容和辅助元数据，以及一个 [约束]，要求评估内容质量、语义相关性、来源权威性和时效性，然后给出最终的搜索相关性类别。目标首先生成逐维度论证 $\mathrm{R}_{q,d}$，随后输出按顺序排列的相关性标签，并以终端六级标签 y(rel) q,d 结束。这种格式明确了终端决策标记的位置，同时保留了用于推导该决策的结构化证据。我们采用提示词驱动的形式，完整的提示词和目标模板见附录 A.1。

4 方法

4.1 LARF 概述

如图 1所示，LARF 遵循三阶段训练到服务流程。第一阶段 **引导式思维链冷启动监督微调**选择具有信息量的查询-文档对，并通过结构化推理依据增强其多维相关性标签，为五个相关性维度提供以推理为导向的监督。第二阶段 **带边际对齐的判别式 GRPO**将逐属性、逐列表和规则一致性奖励与终止词元边际损失相结合，从而将轨迹级对齐与直接的序数决策优化联系起来。第三阶段 **双头知识蒸馏**将推理依据生成作为共享骨干网络的辅助训练信号，并在推理时仅保留多维分类头。三个阶段分别支持结构化相关性学习、终止决策对齐和非自回归在线服务。

4.2 基于引导式思维链的冷启动监督微调

为后续强化学习建立以推理为导向的初始化，我们引入了冷启动监督微调（SFT）阶段。该阶段并非不加选择地扩展训练语料库，而是使用由多智能体合成的思维链（CoT）推理依据，并配以精选的信息丰富实例。值得注意的是，实证评估表明，这种冷启动范式仅使用三**分之**一的数据，便能达到与全数据集调优相当的性能。

多标准策略性数据选择

我们的筛选流程结合了面向覆盖的采样与边界聚焦采样。首先，**分层主题覆盖**会跨垂直领域（如金融、娱乐和医疗保健）进行采样，以拓宽输入分布。其次，**相关性类别密度**会为每个采样查询保留候选文档，使其涵盖尽可能多的不同终止相关性等级。随后，我们将这些样本与来自两个反馈渠道的定向困难样例相结合：实证错误样例，即基线模型与标注的终止标签不一致的样例；以及高熵实例，即基线模型赋予较高预测熵$H\left(\hat{y}_{(\text{相关性})\,q,d} \mid x_{q,d}\right)$的样例，表明模型在推理过程中存在强烈的模型歧义。最终得到的子集既包含普通样例，也包含接近不确定序数边界的实例。

多智能体思维链推理依据合成

为了构建冷启动语料库的结构化推理依据目标，我们采用自动化的多智能体合成流程，通过维度级分析扩充人工相关性标签。该流程通过以下阶段分解每个相关性判断：

1. **维度分解：** 我们部署了一组专门的大语言模型智能体，每个智能体负责评估搜索相关性的一个特定子维度（例如，内容质量、语义相关性、来源权威性和时效性）。2. **独立推理生成：** 每个智能体独立评估 $x_{q,d}$，包括相关的辅助元数据，并生成维度专属分析及中间序数评估。3. **共识协调：** 集中式协调器将各维度专属分析汇总为统一推理依据和暂定总体评估。4. **标签条件化目标构建：** 构建监督微调目标时，协调器以人工标注的标签向量$\mathbf{y}_{q,d}$ 为条件，从而生成$T_{q,d} = (R_{q,d}, \mathbf{y}_{q,d})$。这种条件化仅用于构建标签一致的训练目标；在后续 GRPO 阶段，它不会约束策略生成的推理依据。5. **幻觉检查与重新生成：** 基于提示词的检查器会根据相应的黄金标签，筛查合成的推理依据$R_{q,d}$ 。未通过检查的推理依据将重新生成，多次未通过的则予以丢弃。只有经过验证的目标才会纳入冷启动语料库。

检查器在语料库构建期间离线运行，筛查已根据金标准标签进行条件化的目标。最终，通过将策略性数据选择与这种多智能体思维链合成相结合，我们的流程赋予模型对相关性标准细致入微、多角度的理解。这种稳健的初始化有助于模型在后续强化学习阶段稳定、高效地收敛。

4.3 带间隔对齐的判别式 GRPO

虽然冷启动监督微调奠定了推理基础，但它受限于数据规模，也容易导致模型记忆推理路径。因此，我们引入一个强化学习阶段，在保留对终端决策直接监督的同时，利用多种相关性信号评估采样响应。此外，由于搜索相关性本质上是一项分类任务，模型还必须具备强大的判别能力。

带间隔对齐的判别式 GRPO 将多维奖励系统与施加于末端相关性词元的损失相结合。奖励组件鼓励属性一致性、查询内排序和规则一致性，而间隔项则明确拉开金标准终端标签与其最强竞争标签之间的距离。

多维奖励分解

对于针对 $x_{q,d}$ 的每个采样响应，我们将奖励 $r_{q,d}$ 分解为三个组成部分：属性级奖励、列表级排序奖励和规则一致性奖励。

(1) 属性级相关性奖励（r_{att}）： 仅凭整体相关性标签，无法充分反馈违反了哪项相关性标准。因此，我们使用第 $\mathcal{K}$ 节中定义的3.1 五维标签集，并按如下方式计算属性级奖励：

$$r_{\text{att}} = \sum_{k \in \mathcal{K}} \omega_k(t) \cdot \exp\left(- \left| y_{q,d}^{(k)} - \hat{y}_{q,d}^{(k)} \right| \right). \tag{2}$$

其中，$y_{(k)\,q,d}$ 和$\hat{}y_{(k)\,q,d}$ 分别表示维度k的真实标签和预测序数标签，$\omega_k(t)$ 是动态的维度特定权重。指数项将序数误差转换为平滑且有界的奖励。通过联合优化这些辅助维度，模型能够形成严谨的多视角理解，从而显著提升主要相关性预测的准确率和鲁棒性。

(2) 列表级排序奖励： 标准的大语言模型优化范式通常将搜索相关性视为孤立的点式分类任务。然而，在现实搜索场景中，相对排序顺序——具体而言，即优先排列高度相关的文档而非相关性较低的文档——同样至关重要。为明确优化同一查询q下多个文档的相对偏好顺序，我们设计了查询分组列表级奖励。对于查询组 $\mathcal{G}_q$，令P_d、N_d和T_d 分别表示文档d 相对于所有其他候选文档$d' \in \mathcal{G}_q \setminus \{d\}$的一致对、非一致对和平局对数量。这些数量通过比较预测的终端标签$\hat{}y_{(\text{rel})}$ 与真实终端标签$y_{(\text{rel})}$得出；真实标签相同的成对样本将被排除。我们采用成对归一化相关性（PNR）贡献，其定义为归一化的逐文档一致性得分：

$$c_d = \frac{P_d + 0.5\,T_d}{P_d + N_d + T_d} \in [0, 1]. \tag{3}$$

为明确惩罚排序逆转并强化正确的成对偏好，我们对原始一致率 c_d应用零中心化变换。通过使用$(2c_d - 1) \in [-1, 1]$将 c_d 映射到随机猜测阈值附近，我们建立了严格的基线：仅当策略优于随机排列时，才会激活正向强化（> 0）；而不一致的排序则会受到比例惩罚（最低可达 -1）。为进一步稳定训练，我们引入组规模感知置信权重 $w(n_q)$，根据查询组规模 $n_q = |\mathcal{G}_q|$校准列表级统计量的可靠性，其中 n_q 表示查询组中候选文档的数量，$\bar{n}$ 表示训练集中的平均查询组规模。权重$w(n_q)$会随组规模增大而增加，以反映成对排序统计量的可靠性；同时，平方根缩放可防止大规模组在奖励中占据主导地位。最终，

通过 GRPO算法优化的列表级奖励表示为：

$$w(n_q) = \min\left(1, \sqrt{\frac{n_q - 1}{\bar{n} - 1}}\right), \tag{4}$$

$$r_{\text{list-wise}} = w(n_q) \cdot (2c_d - 1). \tag{5}$$

（3）规则一致性推理奖励（r_{rule}）： 标签条件化SFT目标不足以确保强化学习期间生成的推理依据始终遵循标注原则。因此，我们定义了一个显式规则集 $\mathcal{R}$，其中每条规则都将可验证的相关性条件映射到兼容的维度级或终端标签（例如，过时内容不应获得较高的时效性标签）。对于每个采样展开，二元规则检查器会评估生成的推理依据$R^{\text{gen}}{}_{q、d}$ 以及终端预测$\hat{y}$（rel）$_{q、d}$ 与 $\mathcal{R}$的一致性：

$$r_{\text{rule}} = \begin{cases} +1, & \text{if Consistent}_{\mathcal{R}}\left(R^{\text{gen}}_{q,d}, \hat{y}^{(\text{rel})}_{q,d}\right) = 1, \\ -1, & \text{otherwise.} \end{cases} \tag{6}$$

通过惩罚违反人类标注原则的推理路径，r_{rule} 提升了模型的可解释性和鲁棒性，尤其是在边界情况下。采样响应的最终奖励为

$$r_{q,d} = r_{\text{att}} + r_{\text{list-wise}} + r_{\text{rule}}. \tag{7}$$

面向搜索相关性分类的基于间隔的辅助损失

动机 在以分类为中心的搜索相关性排序中，最终预测严格由单个决策数字标记决定（例如“搜索相关性：3”中的“3”）。尽管这一标记极其重要，传统优化范式却无法为这一关键数字提供直接的标记级监督。具体而言，标准GRPO策略梯度仅作用于轨迹级奖励，因此只能产生稀疏且间接的信号。同样，CoT增强的SFT损失也缺乏所需的细粒度优化，无法显式最大化目标类别的概率，并将其与容易混淆的类别拉开距离。为了解决这一缺陷，我们提出了一种铰链式间隔损失，专门应用于该决策标记的logits。该辅助损失显式扩大真实数字与其最强竞争类别之间的决策边界，同时，其方向性裁剪特性可防止在已经充分分离的样本上过拟合。

公式 设$\mathbf{z} = (z_0, \ldots, z_5) \in \mathbb{R}^6$ 表示在指定的终端标签位置处，限定于六个候选数字标记的 logits，并设$g = y(\text{rel})_{q,d} \in \mathcal{Y}$ 为真实终端相关性标签。我们将经验分类间隔定义为 $\Delta(\mathbf{z}, g)$：

$$\Delta(\mathbf{z}, g) = z_g - \max_{c \neq g} z_c, \tag{8}$$

当且仅当该标记被正确预测时，该值保持严格为正。每个样本的辅助目标采用带目标间隔超参数的铰链式形式 $m > 0$（默认$m = 1.0$）：

$$\mathcal{L}_{\text{margin}}(\mathbf{z}, g) = \left[m - \Delta(\mathbf{z}, g)\right]_+ = \max\Big(0,\ m - \big(z_g - \max_{c \neq g} z_c\big)\Big). \tag{9}$$

对于一个微批次 $\mathcal{B}$，我们将 $\mathcal{B}_{\text{valid}} \subseteq \mathcal{B}$ 定义为这样一个实例子集：其格式化目标恰好包含一个可识别的终端相关性标签位置，且其真实标签属于$\mathcal{Y}$。聚合辅助目标损失取该有效子集上的经验均值：

$$\mathcal{L}^{\mathcal{B}}_{\text{margin}} = \frac{1}{|\mathcal{B}_{\text{valid}}|} \sum_{i \in \mathcal{B}_{\text{valid}}} \left[m - \big(z^{(i)}_{g_i} - \max_{c \neq g_i} z^{(i)}_c\big)\right]_+. \tag{10}$$

与 GRPO 集成 该间隔损失以系数 λ_{margin} 联合纳入主要的 GRPO 策略优化目标：

$$\mathcal{L}_{\text{GRPO+margin}} = \mathcal{L}_{\text{GRPO}} - \beta_{\text{ent}} \mathcal{H}(\pi_\theta) + \lambda_{\text{margin}} \mathcal{L}^{\mathcal{B}}_{\text{margin}}. \tag{11}$$

其中，$\mathcal{L}_{\text{GRPO}}$ 是组相对策略优化损失，其中相对于参考策略 π_{ref} 的 KL 散度惩罚已在奖励函数中进行隐式正则化。$\mathcal{H}(\pi_\theta)$ 表示策略熵，并通过 β_{ent} 进行缩放以鼓励探索；λ_{margin} 则用于平衡终端词元的间隔损失。该辅助项产生的梯度仅通过终止决策位置上的六个数字词元逻辑值传播。

对终止决策的影响。间隔目标通过两种机制改进终止分类：**(i) 自适应边界强调：** $\max^{c\neq g} z_c$ operator 会为每个样本识别最具竞争力的类别。通过迫使优化过程直接扩大最脆弱的决策边界，该方法有效消除易混淆类别之间的歧义，并显著提升边界案例上的准确率。**(ii) 直接终止词元监督：** 通过将该间隔损失专门锚定到最终的搜索相关性类别词元，我们确保在整个强化学习阶段，对这一关键决策点的优化始终处于最高优先级。这提供了直接且高效的监督信号，使正确标签的概率最大化，从而克服轨迹级奖励的稀疏性。

4.4 双头知识蒸馏

为降低工业部署中的自回归思维链延迟，我们进一步训练了一个以 GRPO 训练模型初始化的双头教师模型。该模型在训练期间使用推理依据监督，并在部署时采用仅分类推理，旨在保留强化学习获得的推理能力，同时不生成多步推理依据（Xu 等人，2025年）。

该模型由统一的共享 Transformer骨干网络组成，并配备两个任务专用并行预测层：判别式分类模块 H_{cls} 和生成式推理头 H_{gen}。我们并非仅预测整体相关性，而是将分类模块实例化为五个子头 $\{H^{(k)}_{\text{cls}}\}_{k\in\mathcal{K}}$。每个子头预测 $\mathbf{y}_q$ 的一个组成部分$_d$。联合训练目标为

$$\mathcal{L}_{\text{DHD}} = \lambda_{\text{cls}} \sum_{k\in\mathcal{K}} \mathcal{L}_{\text{CE}}\big(H^{(k)}_{\text{cls}}(\mathbf{h}_n), y^{(k)}_{q,d}\big) + \lambda_{\text{gen}} \mathcal{L}_{\text{LM}}\big(H_{\text{gen}}(\mathbf{H}), R^*_{q,d}\big), \tag{12}$$

其中，$\mathbf{h}_n \in \mathbb{R}^d$ 是最后一个输入词元的隐藏状态，$\mathbf{H}$ 是序列隐藏状态矩阵，而$R^*_{q,d}$ 是经 GRPO 训练的教师模型针对 x_q 生成的推理依据$_d$。系数 λ_{cls}和 λ_{gen} 分别用于平衡分类监督与推理依据监督。生成式推理头是一个仅在训练期间使用的辅助目标，旨在促使共享骨干网络编码推理相关信息。

在将搜索相关性模型蒸馏到用于在线排序的学生模型时，生成式推理头 H_{gen} 会被完全弃用。经过优化的教师模型仅通过多维分类头 H_{cls}执行一次非自回归前向传播。移除该架构后，模型在完全绕过逐词自回归解码所带来的延迟开销的同时，仍能获得具备推理能力的大语言模型所提供的高准确率语义对齐能力。

5 实验

5.1 数据集与基线

为评估我们的框架，我们整理了一个高质量、人工标注的网页搜索相关性基准，其中包括**训练集**（~300K 个样本）和**测试集**（~16K 个样本）。训练数据采用分层标注方案——将主要相关性得分与四个子维度（语义相关性、时效性、内容质量和来源权威性）相结合——以实现稳健的多维监督。同时，测试集经过专门整理，包含困难负例和具有挑战性的边缘案例（例如缺失词元、拼写错误和复杂的语义变体），以严格评估泛化边界。

我们将提出的框架与两种主要基线范式进行比较：(1) 非推理重排序器 Qwen3-Reranker-4B 和 Qwen3-Reranker-8B（Zhang 等人，2025a），这类模型通过单次前向传播直接预测相关性评分，不生成中间推理依据；(2) 基于推理的大语言模型 Qwen3-8B-Instruct 和 Qwen3-235B-A22B-Instruct（Yang 等人，2025年），这类模型先以自回归方式生成推理依据，再预测最终标签。在我们的

模型	细粒度相关性层级				粗粒度对齐		
	NDCG	宏平均 F1	加权 F1	Acc.	精确率	Rec.	F1
Qwen3-Reranker-4B	72.23	34.10	40.80	42.30	61.23	66.93	63.95
Qwen3-Reranker-8B	74.20	36.50	42.30	44.10	64.50	68.10	66.25
Qwen3-1.7B-Base	70.15	31.20	38.50	40.10	58.40	63.50	60.85
Qwen3-8B-Instruct	74.51	37.40	43.15	45.20	68.30	77.20	72.48
Qwen3-235B-A22B-Instruct	75.87	38.10	44.20	45.61	70.50	79.10	74.54
LARF-1.7B已部署的学生模型	75.20	38.05	43.90	45.80	69.10	78.30	73.41
LARF-8B 冷启动监督微调	76.10	40.25	45.30	46.12	71.40	81.15	75.95
LARF-8B 冷启动监督微调 +*GRPO*	78.40	42.15	48.20	49.35	73.15	84.30	78.33
LARF-8B 冷启动监督微调 +*GRPO* +双头蒸馏	79.56	43.68	50.13	50.98	74.82	**86.45**	80.22
LARF-235B 冷启动监督微调 + *GRPO* + 双头蒸馏	**80.03**	**45.76**	**52.39**	**53.23**	**79.99**	85.92	**82.85**

表1：两个语义粒度下的总体离线性能。**DH** 表示双头蒸馏。每列中的最佳值以粗体显示。

框架下使用这些生成式骨干网络进行微调的模型统称为**LARF**。所有 LARF 变体均沿用其对应基线的骨干网络、训练语料库和解码配置，因此报告的差异可归因于该框架，而非模型容量。

5.2 评估指标

在离线评估中，我们报告标准的归一化折损累计增益（NDCG），以衡量列表级排序质量（Wang 等人，2013）。为评估分类性能，我们将搜索相关性映射到六级评分系统：完全不相关（存在安全违规）、不相关、部分相关、基本相关、高度相关和完全相关。基于这一体系，我们引入两项诊断指标：首先，**细粒度相关性层级**衡量所有六个等级的精确匹配准确率、宏平均 F1 和加权 F1，以反映模型对搜索相关性细微变化的敏感度。其次，**粗粒度相关性对齐**将六个等级合并为二元分桶（不相关和相关），用于评估宏观层面的对齐、边界容忍度和总体排序稳定性。

5.3 总体表现

表1 显示，LARF 的主要提升在于区分相邻相关性等级的能力，而非仅仅将明显相关的文档排在更高位置。与其 Qwen3-8B 骨干网络相比，LARF-8B 将细粒度准确率从 45.20% 提高至 50.98%，将 NDCG 从 74.51 提高至 79.56；与表现最强的非推理重排模型 Qwen3-Reranker-8B 相比，其准确率高出 6.88 个百分点，粗粒度 F1 高出 13.97 个百分点。更值得注意的是，LARF-8B 在所有指标上均优于规模近 30 倍的 Qwen3-235B-A22B-Instruct 基础模型，这表明这些提升源于所提出的学习框架，而非模型容量的增加。为取得最佳总体表现，我们采用 235B 模型作为骨干网络，并实现完整的 LARF 框架，取得表中最高成绩：准确率为 53.23%，NDCG 为 80.03。三个阶段的贡献符合其设计预期。引导式思维链冷启动监督微调达到 46.12% 的准确率和 76.10 的 NDCG。这一结果主要归功于对推理过程的监督。随后，带间隔对齐的判别式 GRPO 为 LARF 带来了最大增益：准确率提升 3.23 个百分点，而前一阶段仅提升 0.92 个百分点。这证实了我们的假设：主要难点在于使终端决策与含糊的人工标注边界对齐，而非获得表层语义匹配能力。双头蒸馏又将准确率提高 1.63 个百分点，达到 50.98%，同时从在线路径中移除自回归解码，因此在不牺牲准确率的情况下提升了效率。粗粒度结果呈现出相同的阶段排序，召回率从 77.20 升至 86.45，F1 从 72.48 升至 80.22。

我们注意到，8B 变体保持了最高的粗粒度召回率（86.45），而 235B 变体达到了最高的粗粒度精确率（79.99），因此二者在二分类边界上对两类错误进行了权衡。另一方面，经过蒸馏的 LARF-1.7B 学生模型达到了 45.80% 的准确率和 75.20 的归一化折损累计增益，在参数量约为教师模型五分之一的情况下保留了其 89.8% 的准确率，并且比自身的 1.7B 骨干网络提升了 5.70 个百分点。尤其值得注意的是，这一非自回归学生模型比 8B 基础模型高出 0.60 个百分点，比 Qwen3-Reranker-8B 基线高出 1.70 个百分点，这正是部署变得切实可行的原因：通过引导式思维链和判别式GRPO获得的能力，在压缩到足以在线服务的小型模型后仍得以保留。

模型	细粒度相关性层级			粗粒度相关性对齐		
	宏平均 F1	加权 F1	准确率	精确率	召回率	F1
Qwen3-8B完整数据	37.40	43.15	45.20	68.30	77.20	72.48
LARF-DataSelect 随机	36.85	42.60	44.80	67.90	76.50	71.93
LARF-DataSelect *STC*	38.12	43.50	45.10	69.20	78.10	73.38
LARF-DataSelect *STC + RCD*	39.29	44.23	45.32	70.57	79.23	74.64
LARF-8B *STC + RCD + GCoT*	**40.25**	**45.30**	**46.12**	**71.40**	**81.15**	**75.95**

表2：冷启动监督微调结合引导式思维链的消融研究。**STC**表示分层主题覆盖率，**RCD** 表示相关性类别密度，**GCoT**表示引导式思维链推理依据监督。所有 LARF-DataSelect 行均使用三分之一的训练语料库，而第一行使用完整语料库，作为预算基准。每列中的最佳值以粗体显示。

设置 / 奖励组合	细粒度相关性层级			粗粒度对齐		
	宏平均 F1	加权 F1	准确率	精确率	召回率	F1值
去除 GRPO 第一阶段监督微调	40.25	45.30	46.12	71.40	81.15	75.95
GRPO 属性 + 列表	41.10	46.25	47.18	71.95	82.10	76.69
GRPO 属性 +列表 + 规则	41.23	47.00	47.40	71.28	83.19	76.78
GRPO 属性 + 列表 + 规则 +间隔	**42.15**	**48.20**	**49.35**	**73.15**	**84.30**	**78.33**

表3：判别式GRPO框架中各组件的消融研究。**属性** 表示属性级奖励，**列表** 表示列表级奖励，**规则**表示规则一致性奖励，**间隔** 表示在末端相关性词元上的基于间隔的辅助损失。所有行均从相同的第一阶段监督微调检查点开始。每列中的最佳值以粗体显示。

5.4 消融研究

引导式思维链冷启动监督微调

表 2 分离了第一阶段的两个要素，即样本选择和推理依据监督。在固定为语料库三分之一的数据预算下，与全量数据训练相比，随机采样导致每项指标均有所下降，这证实了数据规模本身具有局限性。分层主题覆盖（记为 STC）在相同预算下将宏平均 F1 从36.85提升至38.12，超过全量数据基线的37.40；而平衡相关性类别密度（记为 RCD）又进一步提升至39.29。因此，所选子集仅使用三分之一的标注数据，宏平均 F1 却比全量数据训练高出1.89个百分点，粗粒度 F1 高出2.16个百分点。这些结果表明，将经过筛选的高质量数据与显式推理相结合，显著优于在大规模、未经精炼的数据集上进行训练。

在所选子集之上加入引导式思维链（GCoT），带来了该阶段最大的一次单项增益。GCoT 行仅使用通过思维链检查器的推理依据，因此报告的增益反映的是经过筛选的合成目标，而非未经筛选的合成目标。宏平均 F1 从39.29提升至40.25，细粒度准确率从45.32%提升至46.12%，粗粒度召回率从79.23提升至81.15。由于最后两行使用的底层样本完全相同，因此这一增益可归因于推理依据监督。

带边际对齐的判别式 GRPO

表3 逐步消融了奖励构成和辅助目标。与第一阶段监督微调模型相比，将属性级奖励与列表级奖励结合后，各项指标均有所提升：细粒度准确率从 46.12% 提高到 47.18%，粗粒度 F1 从 75.95 提高到 76.69。这表明，对相关性属性和组内排序进行分解式监督，能够提供超越单一终端正确性奖励的有效优化信号。

加入规则一致性奖励提升了细粒度质量，同时重新分配了粗粒度错误。宏平均 F1 从 41.10 提升至 41.23，加权 F1 从 46.25 提升至 47.00，准确率从 47.18% 提升至 47.40%。在二分类边界上，召回率从 82.10 升至 83.19，而精确率从 71.95 降至 71.28，因此粗粒度 F1 仅从 76.69 小幅变为 76.78。这种表现

模型	细粒度相关性层级			粗粒度对齐		
	宏平均 F1	加权 F1	准确率	精确率	召回率	F1
无双头 自回归	42.15	48.20	49.35	73.15	84.30	78.33
双头 仅分类头	41.25	47.80	48.95	71.12	83.95	77.00
双头 一维	42.98	49.32	50.65	72.59	85.73	78.61
双头 多维	**43.68**	**50.13**	**50.98**	**74.82**	**86.45**	**80.22**

表4：双头蒸馏设计的消融研究。**AR** 表示在推理时进行自回归推理依据解码，仅分类头 表示完全移除辅助生成头，一维 表示保留单个整体相关性头，而 多维则将其扩展为五个子维度头。所有行使用相同的骨干网络和相同的蒸馏语料库。每列中的最佳值以粗体显示。

模型	HSR	LSR	CS	RS
基线	–	–	–	–
LARF-1.7B	+1.77%	−3.22%	+2.49	+3.63%

表5：部署的 LARF-1.7B 学生模型相较于生产基线的在线 A/B 测试改进。

这与奖励实际约束的内容一致。规则一致性项会奖励生成的推理依据与标注指南在各个相关性维度上的一致性，但不会监督“相关”与“不相关”之间的二元判断。强制维度级一致性会让策略更愿意为满足至少一个维度的边界文档赋予非零等级，从而以牺牲精确率为代价提高召回率，而汇总后的粗粒度 F1 几乎不变。由于我们任务的终端目标是细粒度序数判断，因此我们保留这项奖励，并通过辅助间隔损失处理剩余的边界歧义。

间隔损失带来了第二阶段中最大幅度的提升，也解决了上述精确率下降的问题。细粒度准确率从 47.40% 提升至 49.35%，加权 F1 从 47.00 提升至 48.20，粗粒度精确率从 71.28 回升至 73.15，粗粒度 F1 从 76.78 提升至 78.33。由于该目标会在终端决策标记处明确拉开相邻相关性等级的 logits，因此由此带来的收益主要集中在仅凭奖励信号最难处理的情形，即相邻等级之间的近似平局案例。

双头蒸馏

表4 在固定骨干网络的条件下，对双头设计进行消融。完全移除辅助生成目标并训练纯分类器后，仅分类头这一行的细粒度准确率达到 48.95%，比 49.35% 的自回归推理低 0.40 个百分点。因此，单独使用分类头无法保留推理依据监督提供的全部信号，我们的设计所带来的收益也不能归因于以分类取代生成。重新引入辅助生成头后，这一差距不仅得到弥补，结果也更进一步：在一维设置下，细粒度准确率提升至 50.65%；将分类器扩展为五个子维度头后，则取得表中最佳结果，细粒度准确率达到 50.98%，粗粒度 F1 达到 80.22%。

这一提升源于推理依据监督的表示学习与多维度监督的结合；生成头的作用是在训练时对共享骨干网络起正则化作用，而不是作为推理时执行的组件。此外，子维度监督并非无关紧要的辅助细节：将单一相关性目标改为五个子维度目标后，在训练数据不变的情况下，细粒度准确率提升了 0.33 个百分点，粗粒度 F1 提升了 1.61 个百分点。

总体而言，我们验证了三个阶段中所有组件的积极贡献，证明了每个组件的有效性。

5.5 成本分析

表 6 报告了相同配置的服务成本，因此可以对照查看准确率和成本。自回归推理依据生成对于每个查询-文档对需要耗时 5.68 秒，这对于生产环境中的排序阶段来说难以接受，因为该阶段必须在单次请求预算内对数十个候选项进行评分。在服务时丢弃生成头后，已部署的多维变体耗时降至 0.97 秒，相当于 5.86 倍的加速比和 82.9% 的延迟降低，同时细粒度准确率从 49.35% 提升至 50.98%。

服务路径	解码 g Mode	延迟 y（秒/对）	加速比	Fine Acc.
LARF-1.7B 学生模型	单次遍历	–	–	45.80
不含双头 *AR*	多步	5.68	1.00×	49.35
dh 仅分类头	单次遍历	**0.71**	**8.00**×	48.95
dh 一维	单次遍历	0.82	6.93×	50.65
dh 多维	单次遍历	0.97	5.86×	**50.98**

表6：双头设计的服务成本。延迟按每个查询-文档对进行测量，测试设备为 64 × NVIDIA H20GPUs。每列中的最佳值以粗体显示。

随着分类头数量增加，延迟仅小幅上升 each 0.82 秒，随后又升至 0.97 秒。所有分类头共享一次骨干网络前向传播，每增加一个分类头，仅需对最终隐藏状态执行一次线性投影。与 仅分类头变体相比，多维配置的细粒度准确率提高了 2.03 个百分点，而每对样本仅额外增加 0.26 秒的成本。更重要的是，完整部署模型的延迟比自回归替代方案低 82.9%。因此，新增的分类头以有限的边际计算量恢复了有价值的任务特定信号，从而为总体评分在线排序流水线带来了更好的准确率与延迟平衡。这种更高的预测性能、更低的服务延迟以及更低的计算需求相结合，使多维配置成为生产部署中实用且有价值的运行点。

保留生成头还能维持语言模型的生成能力，并提供纯分类器所不具备的分析接口。生成头在在线排序期间处于禁用状态，因此不会引入任何服务开销；但在离线环境中，可以启用它来开展案例研究、错误分析，以及对单个预测进行诊断。尽管在线服务从不调用它，但它能够提供有价值的定性证据，帮助理解模型行为并追踪失败案例。

5.6 在线性能

为了评估我们框架在真实场景中的影响，我们部署了蒸馏后的 LARF-1.7B 学生模型，并在元宝AI搜索应用上开展了大规模 A/B 测试。

高满意度率 (HSR) 衡量在多大程度上，结果能够充分满足搜索意图；低满意度率 (LSR) 则反映明显不令人满意的结果（Jiang 等人，2015）；HSR 越高、LSR 越低越理想。设 $\mathcal{B}$ 表示经过评估的搜索实例集合，$y_i \in \{0,1,2,3,4,5\}$ 表示分配给实例 i 的满意度等级。我们将 HSR 和 LSR 定义为：

$$\begin{aligned} \mathrm{HSR} &= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{I}[y_i \in \{3,4,5\}], \\ \mathrm{LSR} &= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{I}[y_i \in \{0,1\}]. \end{aligned} \tag{13}$$

闭集分数 (CS) 评估模型在多个质量维度上评估和排列检索到的网页的能力。整个评估流程是一个完全黑盒，无法访问其内部检索或排序机制。这种设置最大限度地减少了模型特定偏好和内部信号的干扰，从而对模型的评估能力提供更客观、可比较的评价。相对满意度 (RS) 衡量检索到的结果在多大程度上满足用户的信息需求。它反映了结果相关性与整体搜索体验之间的关系。由于 RS 主要根据在线用户的行为和反馈得出，因此能够反映真实世界中的用户偏好。

如表5所示，LARF 提升了每项核心业务指标。3 至 5 级的高满意度率反映了查询意图得到排名靠前的结果良好满足的查询所占比例，该指标提升了 1.77%；相对满意度用于量化总体评分

相较于基线系统在用户偏好方面的提升，提高了 3.63%。0 至 1 级的低满意度率下降了 3.22%，内部闭集分数提升了 2.49。
这些结果有两点值得注意。首先，离线和在线结论的方向一致：细粒度序数准确率更高的模型，在在线场景中产生的低满意度曝光也更少，这支持了我们将细粒度层级作为主要离线目标的做法。其次，这些提升来自一个 1.7B 的非自回归学生模型，而非 8B 的教师模型。这表明，纯分类服务路径能够满足实时流量的延迟约束，同时仍能传递推理监督带来的收益。

6 结论与未来工作

本文介绍了 LARF，这是一个将大语言模型推理与工业级搜索相关性估计相结合的三阶段框架。通过整合引导式思维链冷启动监督微调、带间隔对齐的判别式 GRPO，以及双头蒸馏，LARF 提升了细粒度序数相关性准确率，明确了相邻等级之间模糊的决策边界，并从在线服务路径中移除了自回归推理依据解码。在所比较的基线中，LARF 取得了最强的离线结果，并提升了一个大规模商业 AI 搜索系统的用户满意度。未来工作将探索多模态相关性判断、更精简的教师推理依据，以及通过实时用户反馈实现在线强化学习，从而构建自我改进生态系统。

参考文献

Negar Arabzadeh 和 Charles LA Clarke。2025年。基于大语言模型的相关性判断方法基准测试。载于第*48*届国际*ACM SIGIR*信息检索研究与开发大会论文集。3194–3204。

字节跳动。2026年。豆包。https://www.doubao.com/。

Yupeng Chang、Xu Wang、Jindong Wang、Yuan Wu、Linyi Yang、Kaijie Zhu、Hao Chen、Xiaoyuan Yi、Cunxiang Wang、Yidong Wang等。2024年。大语言模型评估综述。*ACM*智能系统与技术汇刊15卷3期（2024年），1–45。

Guo Chen、Junjie Huang、Huaijin Xie、Fei Sun和Tao Jia。2026年。LiR3AG：面向检索增强生成的轻量级重排序推理策略框架。载于*AAAI*人工智能大会论文集，第40卷。30201–30209。

Hongzhan Chen、Siyue Wu、Xiaojun Quan、Rui Wang、Ming Yan 和 Ji Zhang。2023年。MCC-KD：多思维链一致性知识蒸馏。arXiv:2310。14747 [cs.CL] https://arxiv.org/abs/2310.14747

Zheng Chu、Jingchang Chen、Qianglong Chen、Weijiang Yu、Tao He、Haotian Wang、Weihua Peng、Ming Liu、Bing Qin 和 Ting Liu。2023年。思维链推理综述：进展、前沿与未来。*arXiv* 预印本 *arXiv:2309.15402*（2023年）。

Eva Eigner 和 Thorsten H¨andler。2024年。大语言模型辅助决策的决定因素。*arXiv* 预印本 *arXiv:2402.17385*（2024年）。

Guglielmo Faggioli、Laura Dietz、Charles LA Clarke、Gianluca Demartini、Matthias Hagen、Claudia Hauff、Noriko Kando、Evangelos Kanoulas、Martin Potthast、Benno Stein 等。2023年。关于大语言模型用于相关性判断的观点。在*2023*年 *ACM SIGIR* 信息检索理论国际会议论文集。39–50。

Google。2025年。搭载 Gemini 3 的 Google 搜索：迄今为止最智能的搜索。https://blog.google/products-and-platforms/products/search/gemini-3-search-ai-mode/。

Jiaxin Guo、Zewen Chi、Li Dong、Qingxiu Dong、Xun Wu、Shaohan Huang 和 Furu Wei。2026年。奖励推理模型。神经信息处理系统进展38（2026年），150477–150510。

Jiepu Jiang、Ahmed Hassan Awadallah、Xiaolin Shi 和 Ryen W White。2015 年。理解和预测分级搜索满意度。载于《第八届 *ACM* 网络搜索与数据挖掘国际会议论文集》。57–66。

Mikhail V Koroteev。2021 年。BERT：自然语言处理与理解中的应用综述。《*arXiv* 预印本 *arXiv:2103.11943* 》（2021 年）。

Xiaoxi Li、Jiajie Jin、Yujia Zhou、Yuyao Zhang、Peitian Zhang、Yutao Zhu 和 Zhicheng Dou。2025年。从匹配到生成：生成式信息检索综述。*ACM TransactionsonInformation Systems* 43，3（2025年），1–62。

Shih-Yang Liu、Xin Dong、Ximing Lu、Shizhe Diao、Peter Belcak、Mingjie Liu、Min-Hung Chen、Hongxu Yin、Yu-Chiang Frank Wang、Kwang-Ting Cheng 等。2026年。Gdpo：面向多奖励强化学习优化的组奖励解耦归一化策略优化。*arXiv* 预印本 *arXiv:2601.05242* （2026年）。

Wenhan Liu、Xinyu Ma、Weiwei Sun、Yutao Zhu、Yuchen Li、Dawei Yin 和 Zhicheng Dou。2025年。Reasonrank：以强大的推理能力赋能段落排序。*arXiv* 预印本 *arXiv:2508.07050* （2025年）。

Chenji Lu, Zhuo Chen, Hui Zhao, Zhiyuan Zeng, Gang Zhao, Junjie Ren, Ruicong Xu, Haoran Li, Songyan Liu, Pengjie Wang 等。2025a。LORE：用于搜索相关性的大型生成式模型。*arXiv* 预印本 *arXiv:2512.03025* （2025年）。

Xuan Lu、Haohang Huang、Rui Meng、Yaohui Jin、Wenjun Zeng 和 Xiaoyu Shen。2025b。重新思考文档排序中的推理：为何思维链效果有限。*arXiv* 预印本 *arXiv:2510.08985* （2025年）。

Jian Luo、Xuanang Chen、Ben He 和 Le Sun。2024年。Prp-graph：通过图聚合向大语言模型提供成对排序提示，以实现有效的文本重排序。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文）。5766–5776。

Jiaxin Mao、Yiqun Liu、Ke Zhou、Jian-Yun Nie、Jingtao Song、Min Zhang、Shaoping Ma、Jiashen Sun 和 Hengliang Luo。2016年。网络搜索中，相关性何时意味着有用性和用户满意度？载于第*39*届国际 *ACM SIGIR*信息检索研究与发展大会论文集。463–472。

OpenAI。2024年。推出 ChatGPTSearch。https://openai.com/index / 介绍 ChatGPT 搜索/。

Yifan Qiao、Chenyan Xiong、Zhenghao Liu 和 Zhiyuan Liu。2019年。理解 BERT 在排序中的行为。*arXiv* 预印本 *arXiv:1904.07531* （2019年）。

Zhen Qin、Rolf Jagerman、Kai Hui、Honglei Zhuang、Junru Wu、Le Yan、Jiaming Shen、Tianqi Liu、Jialu Liu、Donald Metzler 等。2024年。通过成对排序提示，大语言模型能够有效地充当文本排序器。载于计算语言学协会论文集：*NAACL 2024*。1504–1518。

Matthew Renze 和 Erhan Guven。2024年。大语言模型智能体中的自我反思：对问题解决表现的影响。*arXiv* 预印本 *arXiv:2405.06682* （2024年）。

Saharon Rosset、Ji Zhu 和 Trevor Hastie。2003年。边际最大化损失函数。神经信息处理系统进展 16（2003年）。

Zhihong Shao、Peiyi Wang、QihaoZhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、Y. K. Li、Y. Wu 和 Daya Guo。2024年。DeepSeekMath：突破开放语言模型数学推理能力的极限。arXiv:2402。03300 [cs.CL]https://arxiv.org/abs/2402.03300

Yelong Shen、Xiaodong He、Jianfeng Gao、Li Deng 和 Gr´egoire Mesnil。2014年。使用卷积神经网络学习用于网络搜索的语义表示。载于第*23*届万维网国际会议论文集。373–374。

Tian Tang、Zhixing Tian、Zhenyu Zhu、Chenyang Wang、Haiqing Hu、Guoyu Tang、Lin Liu 和 Sulong Xu。2025年。Lref：一种新颖的基于大语言模型的相关性框架，用于电子商务搜索。载于*Companion Proceedings of the ACM on WebConference2025*。468–475。

腾讯。2026年。元宝：腾讯AI助手。 https://yuanbao.tencent.com/。

Yining Wang、Liwei Wang、Yuanzhi Li、Di He 和 Tie-Yan Liu。2013年。NDCG 类型排序指标的理论分析。载于学习理论会议。PMLR，25–54。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、EdChi、Quoc V Le、Denny Zhou 等。2022年。思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展35（2022年），24824–24837。

Sam Wiseman 和 Alexander M Rush。2016年。将序列到序列学习视为束搜索优化。*arXiv*预印本 *arXiv:1606.02960*（2016年）。

Mengzhou Xia、Sadhika Malladi、Suchin Gururangan、Sanjeev Arora 和 Danqi Chen。2024年。LESS：为有针对性的指令微调筛选高影响力数据。*arXiv* 预印本 *arXiv:2402.04333*（2024年）。

Tian Xie、Zitian Gao、Qingnan Ren、Haoming Luo、Yuqian Hong、Bryan Dai、Joey Zhou、Kai Qiu、Zhirong Wu 和 Chong Luo。2025年。Logic-RL：通过基于规则的强化学习释放大语言模型推理能力。*arXiv* 预印本 *arXiv:2502.14768*（2025年）。

Jillian Xu、Dylan Zhou、Vinay Shukla、Yang Yang、Junrui Ruan、Shuhuai Lin、Wenfei Zou、Yinxiao Liu 和 Karthik Lakshmanan。2025年。双头推理蒸馏：利用仅训练时推理提升分类器准确率。arXiv:2509。21487 [cs.CL] https://arxiv.org/abs/2509.21487

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。2025年。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*（2025年）。

Runming Yang、Taiqiang Wu、Jiahao Wang、Pengfei Hu、Yik-Chung Wu、Ngai Wong 和 Yujiu Yang。2024年。LLM-NEO：面向大语言模型的参数高效知识蒸馏。*arXiv* 预印本 *arXiv:2411.06839*（2024年）。

Edward Yeo、Yuxuan Tong、Morry Niu、Graham Neubig 和 Xiang Yue。2025年。揭秘大语言模型中的长思维链推理。*arXiv* 预印本 *arXiv:2502.03373*（2025年）。

Chunyuan Yuan、Yiming Qiu、Mingming Li、Haiqing Hu、Songlin Wang 和 Sulong Xu。2023年。用于电子商务检索中查询意图分类的多粒度匹配注意力网络。载于 *2023*年*ACM Web*会议论文集。416–420。

Ziyang Zeng、Heming Jing、Jindong Chen、Xiangli Li、Hongyu Liu、Yixuan He、Zhengyu Li、Yige Sun、Zheyong Xie、Yuqing Yang 等。2026年。通过强化学习优化小红书搜索中的生成式排序相关性。载于第*32*届*ACM SIGKDD*知识发现与数据挖掘大会论文集，第*1*卷。2551–2561。

Jinghan Zhang、Xiting Wang、Weijieying Ren、Lu Jiang、Dongjie Wang 和 Kunpeng Liu。2025b。Ratt：用于连贯且正确的大语言模型推理的思维结构。载于 *AAAI*人工智能大会论文集，第39卷。26733–26741。

Kaiyan Zhang、Yuxin Zuo、Bingxiang He、Youbang Sun、Runze Liu、Che Jiang、Yuchen Fan、Kai Tian、Guoli Jia、Pengfei Li 等。2025c。大型推理模型强化学习综述。*arXiv*预印本 *arXiv:2509.08827*（2025年）。

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin 等。2025a。Qwen3嵌入：通过基础模型推进文本嵌入与重排序。*arXiv* 预印本 *arXiv:2506.05176*（2025年）。

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou 和 Ji-Rong Wen。2025年。用于信息检索的大语言模型：一项综述。*ACM Transactions onInformationSystems* 44卷1期（2025年），1–54页。

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang 和 Michael Bendersky。2024a。超越是与否：通过对细粒度相关性标签进行评分，改进零样本大语言模型排序器。载于计算语言学协会北美分会*2024*年会议论文集：人类语言技术（第*2*卷：短篇论文）。358–370页。

Shengyao Zhuang、Xueguang Ma、Bevan Koopman、Jimmy Lin 和 Guido Zuccon。2025年。Rank-r1：通过强化学习增强基于大语言模型的文档重排序器中的推理能力。*arXiv* 预印本 *arXiv:2503.06034*（2025年）。

Shengyao Zhuang、Honglei Zhuang、Bevan Koopman 和 Guido Zuccon。2024b。一种利用大语言模型实现高效且极具效率的零样本排序集合式方法。载于第*47*届 *ACM* 信息检索研究与发展国际会议论文集。38–47。

A 附录

搜索相关性评估提示词

[系统]Y

您是一位搜索结果满意度评估专家。请根据所提供的信息，从四个维度评估搜索结果：**语义相关性**、**时效性**、**来源权威性**、以及**内容质量**。你还应预测**查询维度**并给出总体**满意度** 评分，范围为 0 至 5 分。

评分时，请全面考虑查询类型、时效敏感度和所需权威级别等上下文信息。四个维度的相对重要性可能因查询而异。给出评分后，请详细说明判断依据。

[输入信息]

- **查询：** {查询}
- **文档标题：** {标题}
- **文档内容：** {内容}
- **搜索时间：** {搜索时间}
- **文档发布时间：** {发布时间}

[辅助信息]

- **查询领域/主题：** {主题}
- **发布平台：** {网站}
- **平台名称：** {网站名称}
- **平台领域/主题：** {网站主题}
- **平台级别：** {平台级别}
- **ICP许可证类型：** {ICP许可证}
- **作者姓名：** {作者姓名}
- **作者领域/主题：** {作者主题}
- **作者级别：** {作者级别}

[约束]A

h c
针对文档的语义相关性、时效性、来源权威性和内容质量进行评估。然后根据以上信息确定总体满意度分数。首先输出各维度的评估结果及详细说明，最后给出满意度分数。

评估

图2：用于多维搜索相关性评估的提示词模板。

A.1 提示词

本节介绍用于评估搜索相关性的提示词。如图2所示，模型接收有关查询-文档对的元数据，并根据特定的预定义维度输出评分。

A.2 实现细节

我们采用 Qwen3-235B-A22B-Instruct 作为构建 LARF 的骨干网络。所有训练实验均在 64 × NVIDIA H20 GPU 上进行，批次大小为 4，学习率为 1×10^{-5}。训练序列的最大长度设为 8192 个 token，模型训练 2 个轮次。进行 GRPO训练时，我们使用大小为 8 的展开，并训练 2 个轮次，以降低过拟合风险。推理时，我们使用 vLLM 和指定的服务配置，并将温度设为 0，以确保输出具有确定性。