arXiv:2603.02555v1 [cs.IR] 3 Mar 2026

相关性至关重要：一种用于电子商务查询改写的多任务、多阶段大语言模型方法

Aijun Dai*、Jixiang Zhang†、Haiqing Hu*、Guoyu Tang*、Lin Liu*、Ziguang Cheng*
*京东，北京，中国
†清华大学

摘要——**对于电子商务搜索，用户体验通过用户对返回商品的行为反应来衡量，例如点击率和转化率，以及返回商品与搜索查询之间的相关性。因此，相关性和用户转化构成了查询改写的两个主要目标。查询改写旨在弥合用户表达与商品描述之间的词汇差距。本研究提出了一种以大语言模型为基础的多任务多阶段查询改写框架。值得注意的是，与以往主要侧重于生成改写查询的工作不同，我们将相关性任务融入查询改写。具体而言，该方法利用基于京东用户数据和商品信息预训练的模型，首先进行多任务监督微调（SFT），其中包括改写查询生成任务，以及查询与改写之间的相关性标注任务。随后，我们采用组相对策略优化（GRPO）对模型进行目标对齐，以提升相关性并促进用户转化。通过离线评估和在线 A/B 测试，我们的框架显著提升了电子商务查询改写的有效性，提高了搜索结果的相关性，并增加了每位用户的购买次数（UCVR）。自2025年8月起，我们的方法已在京东这一中国领先的在线购物平台上投入使用。**

关键词—**查询改写、电子商务搜索、大语言模型。**

I. 引言

在信息检索领域，查询改写是一种非常有效的方法，可以弥合用户表达与搜索内容之间的语义差距，帮助用户获取符合其期望的信息。同样，在电子商务搜索中，查询改写通过尽可能减少用户查询与商品描述之间的词汇差异来检索商品，从而更好地满足用户的购物需求并提升整体用户体验。在电子商务搜索中，用户体验主要通过用户与搜索结果的交互以及检索项目与原始查询的相关性来评估。作为电子商务搜索系统的核心组件，查询改写需要同时优化两个关键目标：用户表现和感知相关性。

在查询改写中，主要有两种范式：判别式方法和生成式方法。判别式方法通过检索相似词语来改写查询，通常使用 BM25 和分层可导航小世界图（HNSW）等稀疏或稠密检索技术 [1]。然而，这些方法依赖精确的词汇匹配，因此在语义上较为脆弱，也无法识别不同表达之间的语义等价性。相比之下，生成式方法 [2], [3] 使用语言模型直接将查询改写为不同形式。判别式方法和生成式方法都忽视了相关性在查询改写中的重要性，只关注改写的词汇多样性和用户转化率。在搜索检索漏斗中，查询改写会与其他召回方法竞争共享的检索预算。无节制地增加改写的文本多样性可能会占用全部检索预算，限制通过其他途径检索相关商品，从而显著降低整体召回质量。

近年来，随着大语言模型的迅速发展，越来越多的研究提出了基于大语言模型的查询改写方法。大多数此类工作[4], [5]采用包含面向相关性组件的复合奖励函数。尽管强化学习奖励中包含相关性，但复合公式中的竞争信号往往会削弱其影响，从而限制大语言模型在查询改写过程中以细粒度方式优化相关性的能力。为解决这一局限，我们的方法在融合强化学习奖励的基础上，联合优化相关性标签预测与查询改写，从而推动确定性相关性对齐。

在本研究中，我们提出了一种基于大语言模型的多任务监督微调（SFT）和多任务强化学习（RL）框架，如图1所示，该框架将相关性预测融入查询改写。具体而言，在生成改写的同时，后预训练大语言模型还会生成相应的相关性标签。对于监督微调训练数据的构建，我们使用拒绝采样从在线日志中提取查询-改写对。随后，京东相关性模型根据原始查询与改写检索到的商品之间的相关性计算相关性分数；高于预定义阈值的分数对应相关性标签 1，否则为 0。在监督微调阶段，模型通过任务特定损失联合执行查询改写和相关性标签的下一词元预测：改写生成采用交叉熵，相关性标签预测采用二元交叉熵。在强化学习阶段，除了整合 [4]，中的奖励外，我们还为相关性标签

© 2026年 IEEE。允许个人使用本资料。对于任何其他用途，包括在任何现有或未来媒体中重印或再版本资料以用于广告或促销、创作新的汇编作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护部分，均须事先获得 IEEE 的许可。

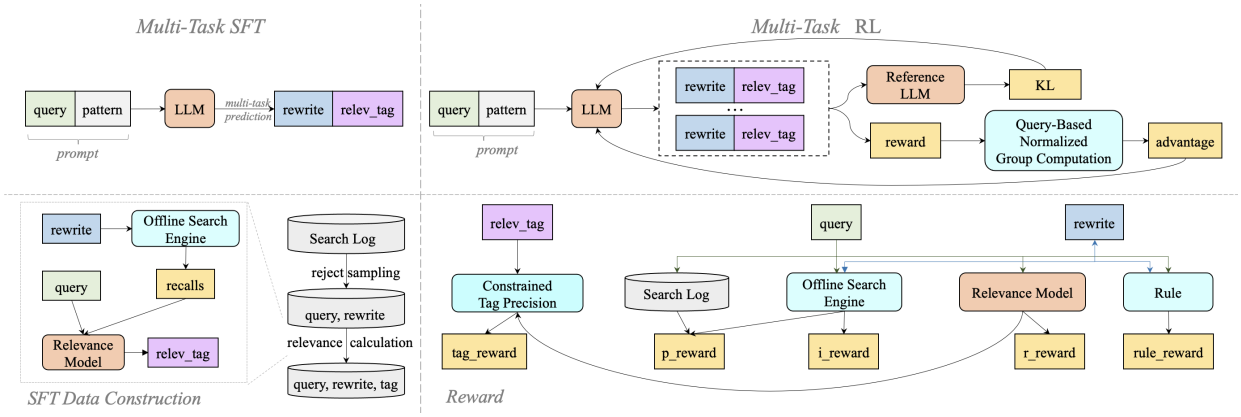

图 1：多任务 SFT 与 RL 架构。(1) 在 SFT 中，大语言模型以提示词作为输入生成改写结果，相关性标签由离线搜索引擎和相关性模型构建。(2) 在 RL 中，大语言模型通过基于查询的归一化 GRPO 进行优化。集成奖励由标签奖励、产出奖励、增量奖励、相关性奖励和规则奖励组成。

预测引入了额外奖励。改写生成的奖励由相关性、生产力、增量和基于规则的组件共同构成。对于相关性标签预测，奖励则根据查询与改写检索到的商品之间的相关性分数得出。

总而言之，本文的主要贡献包括：

- 我们提出了一种多任务多阶段方法，将相关性标签预测融入查询改写中。该方法显著提升了开源大语言模型在电子商务查询改写方面的能力，尤其增强了检索到的商品与用户意图之间的相关性对齐。
- 我们提出了一种强化学习策略，通过优化用于查询改写的复合奖励函数，相较于现有方法，实现了大语言模型在电子商务搜索中的更优对齐。
- 通过全面的离线评估和在线 A/B 测试，我们验证了该框架的有效性，确认了其实际价值和稳健性。

II. 相关工作

A. 查询改写

查询改写在电子商务搜索技术中发挥着关键作用，会影响用户的购物体验以及电子商务平台的收入。总体而言，查询改写方法可分为两类：判别式方法和生成式方法。

判别式方法将查询改写视为检索任务，旨在从候选集中选取恰当的词项，以扩展原始查询的语义。伪相关反馈方法 [6]–[11]通常通过从初次搜索检索到的排名靠前的文档中识别扩展词来增强查询，但可能引入噪声，且难以保持相关性。为缓解这些问题，研究人员探索了利用构建完善的同义词词典 [12], [13] 或挖掘搜索日志 [14]–[18] 来生成候选改写。然而，基于同义词词典的方法严重依赖词典的覆盖范围和质量。基于日志的方法则偏向热门查询，且缺乏足够的数据来处理长尾查询。

生成式方法将查询改写构建为基于 Transformer 的模型所执行的生成过程。零接地和非交互式方法 [19]–[22] 主要依赖大语言模型的先验知识，并通过单次生成来产出改写，缺乏领域适配，导致检索效果有限。仅接地和非交互式方法[23]–[30] 则明确依据语料库证据或伪相关文档进行生成，并通过约束或反馈来控制语义漂移。具备接地能力的交互式方法 [31]–[33] 通过多轮检索—改写循环迭代优化改写，但会增加工程复杂度。

B. 目标对齐

基于人类反馈的强化学习（RLHF）[34] 是使大语言模型与人类偏好对齐的关键技术，已广泛应用于安全保障和推理增强等任务。近端策略优化（PPO） [35] 因其稳定性和有效性，常用于基于人类反馈的强化学习框架。直接偏好优化（DPO） [36]通过利用成对人类偏好数据直接优化模型参数，简化了这一过程。近年来，组相对策略优化（GRPO） [37] 及其变体 [38]–[41]受到广泛关注，并在目标对齐任务中展现出良好效果。对于查询改写，一些方法通过预训练或监督微调（SFT）注入领域专属知识，并进一步利用强化学习算法 [4], [5], [42], [43]对模型目标进行对齐。

III. 方法

A. 框架概述

查询改写旨在增强原始查询的语义，以弥合用户表达与商品描述之间的语义差距，同时确保以相关性为首要考量的用户体验。我们提出了一种两阶段查询

改写框架，由多任务监督微调和由离线搜索引擎驱动的多任务强化学习对齐组成。

B. 多任务监督微调

以往的研究试图通过精心设计的查询改写生成方法，弥合原始查询与其改写之间的词汇差距。然而，这些方法无法保证原始查询与其改写之间的相关性。为应对这一挑战，我们将相关性预测任务融入查询改写，从而确保生成的改写兼具有效性和相关性。

多任务监督微调数据集：我们首先从搜索日志中收集查询-改写对。具体而言，x 是一个查询，改写集合为 $Y=\{y^1,y^2,\cdots,y^n\}$。我们采样数百万条用户查询及其相关改写，以构建初始查询改写数据集 $\mathcal{D}$:

$$\mathcal{D}=\left\{\left(x_i,y_i^k\right)\,\middle|\, x_i\sim p(x),\ y_i^k\sim\pi_{\theta_{\text{old}}}\left(y\mid x=x_i\right),k=1...n\right\} \tag{1}$$

其中，$p\left(x\right)$ 表示我们搜索引擎中的查询分布，$\pi_{\theta_{\text{old}}}$ 表示由 θ_{old}参数化的京东上一代改写策略。

需要注意的是，用户对某个改写所检索到的商品进行的有效点击，是衡量改写 y相关性和有效性的关键指标。因此，我们通过拒绝采样对数据集 $\mathcal{D}$ 应用点击过滤器:

$$\mathcal{D}_r=\left\{\left(x_i,y_i^k\right)\,\middle|\,\left(x_i,y_i^k\right)\in\mathcal{D},\ \mathcal{R}^{y_i^k}\cap\mathcal{C}^{x_i}\neq\emptyset\right\} \tag{2}$$

其中，$\mathcal{R}^{y_i^k}$ 是改写 y_i^k 的召回集，$\mathcal{C}^{x_i}$ 是查询 x_i的点击集。

为了进一步提升生成的改写与原始查询之间的相关性，我们将相关性标签预测纳入查询改写流程。具体而言，根据相关性得分为每个改写分配一个相关性标签:

$$\mathcal{D}_{sft}=\left\{\left(x_i,y_i^k,t)\right)\,\middle|\,\left(x_i,y_i^k\right)\in\mathcal{D}_r\right\} \tag{3}$$

$$t=\begin{cases}1, & \text{if } r_R(x_i,\mathcal{R}^{y_i^k})>\tau^{relev}\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

这里，$r_R(x,\ \mathcal{R}^y)$在公式 6 中定义，τ^{relev} 表示相关性阈值。实验结果表明，提高 τ^{relev} 会提升相关性，但会降低召回率。在本研究中，τ^{relev} 设为 0.2。SFT 训练任务由单任务的改写生成扩展为多任务设置，同时执行改写生成和相关性标注。

监督微调：在我们的多任务监督微调过程中，同时包含改写生成和相关性标注，训练目标是联合最小化改写序列的负对数似然和相关性标签预测的二元交叉熵:

$$\begin{aligned}\mathcal{L}_{\text{SFT}}(\theta)=-\,\mathbb{E}_{(x,y,t)\sim\mathcal{D}_{sft}}\Bigg[&\sum_{i=1}^{|y|}\log\pi_\theta\left(y_i\mid y_{0:i-1},x\right)\\&+\lambda\left(t\log p_\theta(1\mid x,y)+(1-t)\log p_\theta(0\mid x,y)\right)\Bigg]\end{aligned} \tag{5}$$

其中，$(x、\ y、\ t)$分别表示查询、改写结果和相关性标签；$\pi_\theta(\cdot)$表示由 θ参数化的查询改写模型，$p_\theta(t\mid x、\ y)$表示相关性标签的预测概率，而 λ是用于平衡这两个任务的超参数。

C. 多任务强化学习对齐

尽管大语言模型通过 SFT 取得了令人期待的效果，强化学习仍可进一步提升大语言模型的能力 [44] [45]。为了直接优化大语言模型在电子商务领域的查询改写能力，我们采用 GRPO，使大语言模型的目标与提升相关性、促进用户转化保持一致，并利用离线搜索引擎进行对齐。

奖励设计：精心设计的奖励函数对于有效的强化学习至关重要 [46]。我们提出了一种混合奖励，融合搜索引擎反馈和基于规则的组件，以增强多任务目标。

改写奖励：在上下文锚定查询改写中，一个核心挑战是缺乏用于监督训练过程的真实标签。为解决这一问题，我们利用离线搜索引擎的反馈来引导高质量改写，并为其提供奖励，从而优化查询改写。其核心直觉是，高质量的改写不仅应检索出与原始查询相关的商品，还应在检索覆盖方面具备多样性。基于这一原则，我们设计了一种针对改写的基于反馈的奖励，由离线搜索引擎得出的三个部分组成：1）相关性奖励 r_R，用于衡量查询与改写之间的相关性；2）产出奖励 r_P，用于评估改写的有效性；3）增量奖励 r_I，用于量化改写的补充价值。$r_{Feedback}$ 是对这些独立部分的混合组合。

$$r_R(x,\mathcal{R}^y)=\prod_{j=1}^{M}relev(x,\mathcal{R}_j^y) \tag{6}$$

$$r_P(\mathcal{C}^x,R^y)=\sum_{i=1}^{N}\eta(\mathcal{C}_i^x,\mathcal{R}^y)\delta(\mathcal{C}_i^x,\mathcal{R}^y) \tag{7}$$

$$\eta(\mathcal{C}_i^x,\mathcal{R}^y)=\begin{cases}1, & \text{if } \mathcal{C}_i^x\in\mathcal{R}^y\\ 0, & \text{else}\end{cases} \tag{8}$$

$$r_I(\mathcal{R}^x,\mathcal{R}^y)=\frac{|\mathcal{R}^y|-|\mathcal{R}^x\cap\mathcal{R}^y|}{|\mathcal{R}^x|} \tag{9}$$

$$r_{Feedback}(x,y)=r_R(x,\mathcal{R}^y)\times r_I(\mathcal{R}^x,\mathcal{R}^y)+r_P(\mathcal{C}^x,\mathcal{R}^y) \tag{10}$$

其中，x 是查询，y 是改写，相应地，$\mathcal{C}^x$表示查询的点击集，$\mathcal{R}^x$ 是查询的召回集，$\mathcal{R}^y$ 是改写的召回集。$relev(\cdot)$是京东搜索在线相关性模型给出的相关性得分。$|\cdot|$是集合的计数运算。$\delta(\cdot)$是召回位置权重，根据点击的召回位置是否位于前 100%、70% 或 30% 分别设为 1、2 或 3。

此外，为了提升检索多样性，我们引入字符级差异率奖励，用于量化改写相对于原始查询的多样性。此外，我们还加入改写长度约束奖励，以避免输出过长或只是简单地追加词语。这些基于规则的奖励有助于引导模型生成格式良好且可执行的改写：

$$r_{Rule}(x,y) = \alpha \times \text{diff}(x,y) \times \frac{len(x)}{len(y)} \tag{11}$$

其中，α 是用于控制各奖励组成部分之间平衡的超参数。因此，原始查询 x 及其改写 y 的最终改写奖励定义如下：

$$r_{Rewrite}(x,y) = r_{Rule}(x,y) \times r_{Feedback}(x,y) \tag{12}$$

约束标注奖励：我们采用严格的验证机制：对于任何不满足这一结构要求的输出，都将其最终奖励设为零。基于这一约束，若输出符合所需格式，则约束标注奖励定义如下：

$$r_{Tag}(x,y,t) = \begin{cases} 1, & \text{if } |t - \mathbb{I}(r_R(x,\mathcal{R}^y) > \tau^{\text{relev}})| = 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

其中，$\mathbb{I}(\cdot)$表示指示函数，$r_R(x,\ \mathcal{R}^y)$在公式 6 中定义，τ^{relev} 表示相关性阈值。

最终融合奖励是一种混合信号，它结合了搜索引擎反馈和基于规则的组件，以强化多任务目标：

$$r_{Fusion}(x,y,t) = r_{Rewrite}(x,y) \times r_{Tag}(x,y,t) \tag{14}$$

组相对策略优化（GRPO）：鉴于查询改写具有一对多的特性，即同一查询可能存在多个有效改写，我们采用 GRPO，利用候选改写之间的相对奖励对模型进行微调。对于每个查询 x，策略 $\pi_\theta(\cdot)$会采样 N个候选项，记为$\hat{}\boldsymbol{Y} = \{\hat{}y^1,\ .....,\ \hat{}y^N\}$，其中$\hat{}y^i$表示与相应相关性标签配对的改写$y^i$。值得注意的是，此阶段采用束搜索采样，而非 top-k 或 top-p 采样。每个候选项$\hat{}\ y^i$ 的奖励 r^i 根据原始查询 x 计算，定义见公式 14。随后，候选项$\hat{}y^i$ 的组相对优势计算如下：

$$A^i = \frac{r^i - mean(r^1,\ldots,r^N)}{std(r^1,\ldots,r^N)} \tag{15}$$

$\hat{}$ 由于每个改写 y^i 都被序列化为单一输出序列，生成该改写的似然即为模型下的序列似然。概率比定义为：

$$\rho^i = \frac{\pi_\theta(\hat{y}^i|x)}{\pi_{\theta_{\text{old}}}(\hat{y}_i|x)} \tag{16}$$

其中，分子和分母均按整个序列 $\hat{}y^i$ 上的词元级概率乘积计算。最后，GRPO 目标将裁剪策略优化与 KL正则项相结合：

$$\begin{aligned}\mathcal{L}_{\text{GRPO}}(\theta) = & -\mathbb{E}\big[x \sim P(x),\ \{\hat{y}^i\}_{i=1}^N \sim \pi_{\text{old}}(\hat{y}|x)\big] \\ & \frac{1}{N}\sum_{i=1}^{N}\Bigg(\min\left(\rho^i A^i,\ \text{clip}\left(\rho^i,\ 1-\epsilon,\ 1+\epsilon\right)A^i\right) \\ & -\beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Bigg)\end{aligned} \tag{17}$$

其中，KL散度使用无偏估计器进行估计：

$$\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(\hat{y}^i|x)}{\pi_\theta(\hat{y}^i|x)} - \log\frac{\pi_{\text{ref}}(\hat{y}^i|x)}{\pi_\theta(\hat{y}^i|x)} - 1 \tag{18}$$

在该目标中，ϵ指定用于稳定更新的裁剪范围，而 β 控制KL正则化的强度。

D. 在线服务

在将大语言模型部署用于在线服务时，必须同时考虑计算效率与延迟。我们采用近线部署策略 [47] ，以满足京东搜索系统的延迟要求，同时充分发挥大语言模型的能力。对于每个新查询，我们的查询改写大语言模型使用束搜索采样生成一组改写及其对应的相关性标签。在基于相关性标签进行过滤后，选取排名前三的改写作为最终改写。改写生成后，会被缓存最多14天，以实现快速的在线响应。

IV. 实验

A. 数据集

多任务监督微调数据集：如第 III-B 节所述，我们获得了一个多任务监督微调数据集，其中包含约 560 万条查询及其改写，每条数据均标注了相应的相关性标签，表 I 对此进行了展示。

多任务强化学习数据集：该强化学习数据集包含 3.6 万条查询及其对应的点击产品集合，它是根据我们算法中定义的强化学习奖励，利用监督微调数据集构建而成的。

评估数据集：对于离线测试集，我们构建了一个包含 1,000 组 <查询、改写> 配对的人工标注数据集 I，作为评估相关性标注准确率的真实标注。由于改写通过倒排索引方法检索商品，我们采用基于嵌入的检索策略构建了一个召回率评估数据集。最终的评估数据集包含 10 万条查询及其对应的检索商品对，其中的商品是用户针对给定查询频繁点击的商品。

B. 评估

离线指标：为了全面评估我们的查询改写方法，我们采用了多种离线指标：

1) **标注准确率：**一种以准确率为导向的指标，用于衡量我们的LLM生成的相关性标签与经验丰富的标注人员分配的真实标注之间的一致性。

表 I：提示示例和数据示例。对于监督微调，蓝色部分为可选内容。对于强化学习，“查询 {的同义搜索词及其对应的相关性标签}”是大语言模型生成改写和相关性标签的输入。

阶段	提示示例 / 数据示例
SFT	其对应的相关性标签 的相关性标签 {查询} 为 {改写} $<|sep|>$ {相关性标签}
	同义搜索词及其对应的 Kappa 书包相关性标签是 Kappa 双肩包 $<|sep|>$ 1。
RL	查询、商品标题、点击商品 *ID*
	男士香水， {男士 50 毫升淡香水 东方香调，50 毫升茶淡香水 木质香调}， $\{id_1, id_2\}$
评估	查询，商品标题
	婴儿酸奶，SIMPLELOVE 父爱Formula儿童酸奶0%蔗糖原味Flavor 100g*6袋

2) **召回率：**该指标用于评估大语言模型生成的改写对搜索引擎的帮助程度，定义如下：

$$\text{recall} = \frac{\sum_{x \in \mathcal{D}_{recall}} |\mathcal{R}^{\mathcal{Y}} \cap \mathcal{C}^{x}|}{|\mathcal{D}_{recall}|} \tag{19}$$

其中， $\mathcal{D}_{recall}$ 表示召回率评估数据集， $\mathcal{Y}$ 表示按相关性标签筛选出的 x 改写集合， $\mathcal{R}^{\mathcal{Y}}$表示通过倒排索引方法获得的 $\mathcal{Y}$ 召回集， $\mathcal{C}^{x}$ 表示通过基于物品或基于嵌入的检索获取的有效点击集合。 $|\cdot|$ 表示集合的计数操作。

3) **相关性得分：**一种用于衡量改写与查询之间平均相关性得分的语义对齐指标。每个 <查询与改写> 对的相关性得分按公式（6）所述计算。

在线指标：为评估模型的在线表现，我们采用三项指标：用户转化率（UCVR）、用户点击率（UCTR）和独立访客价值（UV价值）。

C. 实现细节

监督微调和强化学习训练过程均使用开源 TRL 框架 [48] 完成。下文将详细说明关键训练参数；所有未提及的参数均保持默认值。

多任务微调：我们使用余弦学习率调度对 Qwen3-8B 模型进行微调，峰值学习率为 2e-5，预热比例为 0.03。优化器采用 AdamW（$\beta_1 = 0.9$、 $\beta_2 = 0.999$），权重衰减为 0.001。批量大小设为 64，并采用 DeepSpeed ZeRO Stage 1 并行策略和 bfloat16 精度。模型在 8 个昇腾 910B NPU 上训练 5 个训练轮次，总耗时 25 小时。

训练后对齐：对齐阶段采用线性学习率调度，学习率设为 1e-6。我们
将 AdamW优化器的系数配置为（β_1， β_2）（0.9，0.999），并省略权重衰减以保持训练一致性。GRPO 中的 KL系数 β 设为 0.4。训练使用 4 块英伟达H800 GPU，持续 18 小时。

D. 离线实验

1）不同大语言模型的比较：本节对基于多任务监督微调数据集训练的不同大语言模型进行比较分析，结果汇总于表 II。我们考察了若干开源大语言模型，包括百川 [49], Qwen2.5 [50], 混元 [51], Qwen3 [52]和 R1蒸馏 Qwen3。其中，R1蒸馏版Qwen3-8B的整体表现优于其他模型。不过，Qwen3-8B的相关性得分最高，而这正是本文的主要关注点。因此，我们选择Qwen3-8B作为基础模型，并在后续实验之前对其进行电商领域预训练。

表 II：基于多任务监督微调数据集训练的不同大语言模型比较。

模型	准确率（%）	召回率(%)	相关性(%)
Baichuan-7B	79.9	62.69	73.03
Qwen2.5-7B	80.5	62.79	72.90
Hunyuan-7B	80.7	62.85	71.57
Qwen3-8B	80.8	62.86	73.21
R1-Qwen3-8B	81.6	63.18	71.79

2) 多任务监督微调与单任务监督微调：鉴于多任务监督微调的训练目标是在单个步骤中同时生成改写及其相关性标签，我们旨在将此方法与单任务监督微调设置进行比较，在后者中，模型仅接受生成改写的训练。

对于单任务设置，我们使用第 III-B 节所述通过拒绝采样获得的 <查询、改写>对进行监督微调，模型仅针对改写生成进行训练。图 2 显示，与采用单任务监督微调的模型相比，多任务监督微调显著提高了相关性，但召回率较低。这些结果凸显了多任务监督微调在提升改写相关性方面的有效性。

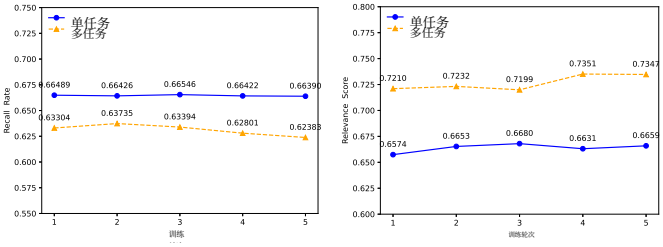

图 2：召回率与相关性得分对比

3）强化学习算法的全面研究：近期进展 [5], [42], [43] 表明，强化学习算法对提升查询改写任务的性能有显著贡献。基于这些发现，我们的对比实验主要聚焦于强化学习阶段，在其中加入了 DPO 和 PPO 基线，并详细分析了所提出框架中各个奖励组成部分的单独作用及其

组合效果。我们使用监督微调模型，通过束搜索从强化学习数据集中为每个查询采样 10 个候选改写，并用融合奖励对其打分，由此构建了包含 72k 个 DPO 训练对的数据集。对于每个查询，我们将得分最高的改写选为正样本，并从得分最低的五个候选改写中随机选取两个负样本。在 PPO 实验中，我们直接使用本文定义的融合奖励作为训练奖励。

表 III：强化学习算法的综合研究

	方法	准确率(%)	召回率(%)	相关性(%)
	预训练+监督微调	80.0	63.74	72.32
预训练+SFT	+DPO	81.2	63.93	72.61
	+PPO	80.9	65.89	70.03
	+GRPO (r_P)	77.2	64.04	74.17
	+GRPO (r_I)	81.6	47.22	33.97
	+GRPO (r_R)	77.4	34.07	84.15
	+GRPO ($r_{Feedback}$)	77.8	62.39	68.98
	+GRPO ($r_{Rewrite}$)	80.1	70.18	64.16
	+GRPO ($r_{Feedback\&Tag}$)	81.9	59.36	73.29
	+GRPO (r_{Fusion})	81.5	67.94	69.65

如表 III 所示，DPO 和 PPO 在整体指标上均带来了中等幅度的提升。对于 GRPO，每个单独的奖励组件都能有效实现其预期目的，但往往会牺牲其他性能指标。融合奖励取得了最佳的整体性能。($r_{Feedback\&Tag} = r_{Feedback} \times r_{Tag}$)

4) 束大小消融研究（*GRPO*）：如上所述，经过多任务监督微调后，我们的模型相关性得分更高，但召回率更低。为了在保持相关性的同时提高召回率，我们采用 GRPO 进行目标对齐。值得注意的是，在此阶段采用的是束搜索采样，而不是 top-k 或 top-p 采样。表 IV 展示了目标对齐阶段不同束大小对模型性能的影响。可以观察到，随着束大小增加，召回率和标注准确率均持续提升，而相关性得分则先升高后下降。

表 IV：束大小消融研究

方法	束大小	准确率(%)	召回率(%)	相关性(%)
预训练+监督微调	/	80.0	63.74	72.32
预训练+监督微调+GRPO	3	62.9	66.76	65.86
预训练+监督微调+GRPO	5	68.7	67.30	66.05
预训练+监督微调+GRPO	**10**	78.5	67.94	69.65
预训练+监督微调+GRPO	15	78.9	68.97	67.79

5）不同重写次数的结果：每个查询经过相关性标签筛选后保留的重写数量也会影响检索性能。显然，增加重写候选项的数量能够显著提高召回率。

表 V：不同重写次数的结果。

重写次数	召回率(%)	相关性(%)
3	67.94	69.65
4	69.44	69.65
5	69.96	69.80

E. 在线实验

为了评估我们模型的线上表现，我们将其部署在在线搜索引擎中，并分配了 10% 的用户流量进行为期七天的 A/B 测试。评估聚焦于三项关键指标：独立访客价值（UV价值）、用户转化率（UCVR）和用户点击率（UCTR）。如表 VI 所示，与上一代改写模型相比，我们的模型在 UV、UCVR 和 UCTR 方面分别提升了 0.106%、0.181% 和 0.082%。这些结果表明，我们的模型能够有效改写查询，并解决语义匹配过程中可能存在的语义差距。

表 VI：在线实验结果。

在线流量	UV价值	uCVR	uCTR
All	+0.106%	+0.181%	+0.082%
Top	+0.240%	+0.170%	+0.110%
躯干	+0.430%	+0.030%	+0.020%
长尾	+0.490%	+0.290%	+0.110%

F. 案例研究

表VII展示了我们模型的一些结果示例。案例1和案例2表明，我们的多任务框架提升了大语言模型在电子商务领域的综合能力，其中“NB”是“新百伦”的缩写。案例3说明，在查询改写中引入相关性任务能够有效减少无关商品的检索。

表 VII：案例样本。

	查询	改写	相关性标签
1	厨房排水管	水槽排水管	1
2	NB羽绒服	新百伦羽绒服	1
3	小米平板	平板电脑	0

五、结论

本文揭示了电子商务查询改写中的一项关键局限：系统性地忽视相关性会严重影响检索质量。因此，我们提出一种多任务多阶段大语言模型框架，将相关性标签预测融入查询改写。在监督微调阶段，大语言模型通过任务特定损失，同时生成改写结果和相应的相关性标签。为使目标与相关性和用户转化相一致，我们采用 GRPO，并通过基于查询的归一化组优化进行优势计算，同时使用综合奖励，其中包括相关性奖励、产出奖励、增量奖励、规则奖励和相关性标签奖励。离线和在线实验验证了我们提出的多任务多阶段框架的有效性。

参考文献

[1] Y. A. Malkov 和 D. A. Yashunin，《使用分层可导航小世界图进行高效且稳健的近似最近邻搜索》，2018 [ ] 1603 09320 。在线。可访问：https://arxiv.org/abs/ 。

[2] Y. Qiu、K. Zhang、H. Zhang、S. Wang、S. Xu、Y. Xiao、B. Long 和 W.-Y. Yang，《通过循环一致性翻译实现电子商务搜索的查询改写》，载于 2021 IEEE 第 37 届国际数据工程会议（ICDE）。IEEE，2021 年 4 月，第 2435–2446 页。 [在线]。可访问：http://dx.doi.org/10 1109 51399 2021 00276 。/ICDE。。

[3] M.-C. Lee、B. Gao 和 R. Zhang，《搜索广告中通过生成对抗网络进行罕见查询扩展》，载于第 24 届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，系列：KDD '18。美国纽约州纽约：计算机协会，2018 年，第 500–508 页。 [在线]。可访问：https://doi.org/10 1145 3219819 3219850 。/ 。

[4] A. Dai、Z. Zhu、H. Hu、G. Tang、L. Liu 和 S. Xu，《增强电子商务查询改写：一种结合领域特定预训练与强化学习的大语言模型方法》，载于第 33 届 ACM信息与知识管理国际会议论文集，2024 年，第 4439–4445 页。[5] W. Peng、G. Li、Y. Jiang、Z. Wang、D. Ou、X. Zeng、D. Xu、T. Xu 和 E. Chen，《基于大语言模型的淘宝搜索长尾查询改写》，载于 ACM万维网大会 2024 伴随论文集，2024 年，第 20–28 页。[6] J. Xu 和 W. B. Croft，《利用局部和全局文档分析进行查询扩展》，载于 ACM SIGIR论坛，第 51 卷，第 2 期。ACM，纽约州纽约，美国，2017 年，第 168–175 页。[7] S. Naseri、J. Dalton、A. Yates 和 J. Allan，《CEQE到SQET：上下文化嵌入用于查询扩展的研究》，《信息检索》，第 25 卷，第 2 期，第 184–208 页，2022 年 6 月。 [在线]。可访问：https://doi.org/10.1007/s10791-022-09405-y [8] ——《CEQE到SQET：上下文化嵌入用于查询扩展的研究》，《信息检索杂志》，第 25 卷，2022 年，第 208 页，第 2 期，第 184 页。[9] Z. Zheng、K. Hui、B. He、X. Han、L. Sun 和 A. Yates，《BERT-QE：用于文档重排序的上下文化查询扩展》，arXiv 预印本 arXiv:2009.07258，2020 年。[10] A. Khader 和 F. Ensan，《面向新冠肺炎学术搜索的查询扩展词排序学习》，《生物医学信息学杂志》，第 142 卷，第 104386 页，2023 年。[11] E. Bassani、N. Tonellotto 和 G. Pasi，《结合上下文化词嵌入的个性化查询扩展》，《ACM信息系统汇刊》，第 42 卷，第 2 期，第 1–35 页，2023 年。[12] J. Bhogal、A. MacFarlane 和 P. Smith，《基于本体的查询扩展综述》，《信息处理与管理》，第 43 卷，第 4 期，第 866–886 页，2007 年。[13] A. Mandal、I. K. Khan 和 P. S. Kumar，《利用自动同义词提取进行电子商务搜索查询改写》。载于 eCOM@ SIGIR，2019 年。[14] I. Antonellis、H. Garcia-Molina 和 C.-C. Chang，《Simrank++ 通过点击图链接分析进行查询改写（海报）》，载于第 17 届万维网国际会议论文集，2008 年，第 1177–1178 页。[15] H. Cui、J.-R. Wen、J.-Y. Nie 和 W.-Y. Ma，《利用查询日志进行概率查询扩展》，载于第 11 届万维网国际会议论文集，2002 年，第 325–332 页。[16] Z. Kunpeng、W. Xiaolong 和 L. Yuanchao，《一种基于查询日志挖掘的新型查询扩展方法》，《亚洲语言处理国际期刊》，第 19 卷，第 1 期，第 1–12 页，2009 年。[17] S. Li、F. Lv、T. Jin、G. Li、Y. Zheng、T. Zhuang、Q. Liu、X. Zeng、J. Kwok 和 Q. Ma，《淘宝搜索中的查询改写》，载于第 31 届 ACM信息与知识管理国际会议论文集，2022 年，第 3262–3271 页。[18] S. Manchanda、M. Sharma 和 G. Karypis，《电子商务查询中的意图词加权》，载于第 28 届 ACM信息与知识管理国际会议论文集，2019 年，第 2345–2348 页。[19] L. Wang、N. Yang 和 F. Wei，《Query2doc：利用大语言模型进行查询扩展》，arXiv 预印本 arXiv:2303.07678，2023 年。[20] R. Jagerman、H. Zhuang、Z. Qin、X. Wang 和 M. Bendersky，《通过提示大语言模型进行查询扩展》，arXiv 预印本 arXiv:2305.03653，2023 年。

[21] L. Gao、X. Ma、J. Lin 和 J. Callan，“无相关性标签的精确零样本稠密检索”，载于《计算语言学协会第61届年会论文集》（第1卷：长篇论文），2023年，第1762–1777页。[22] L. Liu 和 M. Zhang，“Exp4fuse：利用基于大语言模型的查询扩展增强稀疏检索的排序融合框架”，arXiv 预印本 arXiv:2506.04760，2025年。[23] P. Jia、Y. Liu、X. Zhao、X. Li、C. Hao、S. Wang 和 D. Yin，“Mill：利用大语言模型进行零样本查询扩展的相互验证”，arXiv 预印本 arXiv:2310.19056，2023年。[24] X. Chen、X. Chen、B. He、T. Wen 和 L. Sun，“分析、生成与改进：利用大语言模型进行零样本开放域问答查询扩展”，载于《计算语言学协会 ACL 2024 研究成果》，2024年，第11 908–11 922页。[25] Y.-S. Chuang、W. Fang、S.-W. Li、W.-t. Yih 和 J. Glass，“扩展、重排序与检索：开放域问答的查询重排序”，arXiv 预印本 arXiv:2305.17080，2023年。[26] X. Wang、S. MacAvaney、C. Macdonald 和 I. Ounis，“用于有效即席搜索的生成式查询改写”，arXiv 预印本 arXiv:2308.00415，2023年。[27] Y. Lei、Y. Cao、T. Zhou、T. Shen 和 A. Yates，“利用大语言模型进行语料库引导的查询扩展”，arXiv 预印本 arXiv:2402 18031，2024年。[28] L. Zhang、Y. Wu、Q. Yang 和 J.-Y. Nie，“探索利用大语言模型进行查询扩展的最佳实践”，arXiv 预印本 arXiv:2401.06311，2024年。[29] H. Li、X. Wang、B. Koopman 和 G. Zuccon，“伪相关反馈足以弥合小型与大型稠密检索模型之间的差距”，arXiv 预印本 arXiv:2503.14887，2025年。[30] T. Jaenich、G. McDonald 和 I. Ounis，“利用生成式查询扩展实现公平曝光分配”，载于《欧洲信息检索会议》。施普林格，2025年，第267–281页。[31] J. Feng、C. Tao、X. Geng、T. Shen、C. Xu、G. Long、D. Zhao 和 D. Jiang，“搜索与大语言模型在信息检索中的协同作用”，arXiv 预印本 arXiv:2305 07402，2023年。[32] M. S. Rashid、J. A. Meem、Y. Dong 和 V. Hristidis，“面向成本受限数据源检索的渐进式查询扩展”，arXiv 预印本 arXiv:2406.07136，2024年。[33] Y. Lei、T. Shen 和 A. Yates，“Thinkqe：通过演进式思考过程进行查询扩展”，arXiv 预印本 arXiv:2506.09260，2025年。[34] L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、K. Slama、A. Ray 等，“通过人类反馈训练语言模型遵循指令”，《神经信息处理系统进展》，第35卷，第27 730–27 744页，2022年。[35] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，“近端策略优化算法”，arXiv 预印本 arXiv:1707 06347，2017年。[36] R. Rafailov、A. Sharma、E. Mitchell、C. D. Manning、S. Ermon 和 C. Finn，“直接偏好优化：你的语言模型其实暗中充当奖励模型”，《神经信息处理系统进展》，第36卷，第53 728–53 741页，2023年。[37] Z. Shao、P. Wang、Q. Zhu、R. Xu、J. Song、X. Bi、H. Zhang、M. Zhang、Y. Li、Y. Wu 等，“Deepseekmath：推动开放语言模型数学推理能力的极限”，arXiv 预印本 arXiv:2402 03300，2024年。[38] Q. Yu、Z. Zhang、R. Zhu、Y. Yuan、X. Zuo、Y. Yue、W. Dai、T. Fan、G. Liu、L. Liu 等，“Dapo：大规模开源大语言模型强化学习系统”，arXiv 预印本 arXiv:2503.14476，2025年。[39] Z. Liu、C. Chen、W. Li、P. Qi、T. Pang、C. Du、W. S. Lee 和 M. Lin，“理解类 R1-Zero 训练：批判性视角”，arXiv 预印本 arXiv:2503.20783，2025年。[40] Q. Zhang、H. Wu、C. Zhang、P. Zhao 和 Y. Bian，“问对问题，答案就成功了一半：完全无监督的大语言模型推理激励”，arXiv 预印本 arXiv:2504.05812，2025年。[41] C. Zheng、S. Liu、M. Li、X.-H. Chen、B. Yu、C. Gao、K. Dang、Y. Liu、R. Men、A. Yang 等，“组序列策略优化”，arXiv 预印本 arXiv:2507.18071，2025年。[42] B. Zuo、X. Zhang、F. Li、P. Wang、J. Xu 和 B. Zheng，“Value：通过加权字典树在赞助搜索中实现查询改写的价值感知大语言模型”，arXiv 预印本 arXiv:2504.05321，2025年。[43] Y. Zhang、B. Liu、X. Zhang、A. Mandal、Z. Wu 和 C. Xu，“Leser：通过搜索引擎反馈强化学习扩展电商查询”，arXiv 预印本 arXiv:2509.05570，2025年。

[44] D. Guo、D. Yang、H. Zhang、J. Song、R. Zhang、R. Xu、Q. Zhu、S. Ma、P. Wang、X. Bi 等，“Deepseek-R1：通过强化学习激励大语言模型的推理能力”，*arXiv* 预印本 *arXiv:2501 12948*，2025年。[45] T. Chu、Y. Zhai、J. Yang、S. Tong、S. Xie、D. Schuurmans、Q. V. Le、S. Levine 和 Y. Ma，“监督微调会记忆，强化学习能泛化：基础模型后训练的比较研究”，*arXiv* 预印本 *arXiv:2501.17161*，2025年。[46] Y. Gao、C. M. Meyer、M. Mesgar 和 I. Gurevych，“用于高效强化学习的抽取式文档摘要奖励学习”，*arXiv* 预印本 *arXiv:1907.12894*，2019年。[47] X. Amatriain 和 J. Basilico，“工业界的推荐系统：Netflix 案例研究”，载于《推荐系统手册》。施普林格，2015年，第385–419页。[48] L. von Werra、Y. Belkada、L. Tunstall、E. Beeching、T. Thrush、N. Lam-bert 和 S. Huang。（2020）TRL：Transformer强化学习。https://github.com/huggingface/trl。[49] A. Yang、B. Xiao、B. Wang、B. Zhang、C. Bian、C. Yin、C. Lv、D. Pan、D. Wang、D. Yan 等，“百川 2：开放式大规模语言模型”，*arXiv* 预印本 *arXiv:2309.10305*，2023年。[50] Q. Team 等，“Qwen2技术报告”，*arXiv* 预印本 *arXiv:2407.10671*，第2卷，第3页，2024年。[51] 腾讯，“Hunyuan-7B”，https://huggingface.co/tencent/Hunyuan-7B，2025, [在线；访问日期：2025-10-28]。[52] A. Yang、A. Li、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Gao、C. Huang、C. Lv 等，“Qwen3技术报告”，*arXiv* 预印本 *arXiv:2505.09388*，2025年。

.