arXiv:2601.03211v1 [cs.IR] 6 Jan 2026

微调小语言模型，打造高效的企业搜索相关性标注器

Yue Kang[1*†] **Zhuoyi Huang**[1*] **Benji Schussheim**[1] **Diana Licon**[1] **Dina Atia**[1]**Shixing Cao**[1] **Jacob Danovitch**[1†] **Kunho Kim**[1] **Billy Norcilien**[1] **JonahKarpman**[1]**Mahmound Sayed**[1] **Mike Taylor**[1] **Tao Sun**[1] **Pavel Metrikov**[1] **Vipul Agarwal**[2‡]**Chris Quirk**[1] **Ye-Yi Wang**[1] **NickCraswell**[1] **Irene Shaffer**[1] **Tianwei Chen**[1†]**Sulaiman Vesal**[1†] **SoundarSrinivasan**[1†]微软 [2]亚马逊

摘要

在企业搜索中，由于难以获取标注数据，如何大规模构建高质量数据集仍是一项核心挑战。为解决这一问题，我们提出一种高效方法，通过微调小语言模型（SLM）实现准确的相关性标注，其吞吐量高，且在特定领域的标注质量可与最先进的大语言模型（LLM）相当，甚至更胜一筹。为弥补企业领域缺乏高质量且易获取的数据集这一不足，我们的方法利用合成数据生成。具体而言，我们使用大语言模型根据种子文档合成逼真的企业查询，再用 BM25 检索困难负例，并由一个教师大语言模型分配相关性分数。随后将生成的数据集蒸馏到小语言模型中，得到一个紧凑的相关性标注器。我们在一个高质量基准测试集上评估该方法，该测试集包含 923 个由受过培训的人工标注员标注的企业查询-文档对。结果表明，蒸馏后的小语言模型与人工判断的一致性达到或超过教师大语言模型。此外，我们微调后的标注器大幅提升了吞吐量，实现了 17× 的提升，同时也 19× 更具成本效益。该方法支持面向企业级检索应用的大规模、高性价比相关性标注，有助于在真实场景中快速开展离线评估和迭代。

1 引言

随着人工智能驱动的搜索技术不断发展，大语言模型（LLM）在搜索和相关性标注方面展现出了卓越的能力 [1, 2]。然而，企业搜索面临着独特的挑战，这些挑战超越了传统网络搜索的范畴 [3]。企业搜索是指从组织的异构内部存储库中检索信息，例如电子邮件、聊天记录、文档和知识库；用户既会发起基于关键词的查询（例如按文件名或作者搜索），也会发起语义查询（例如就文档提出自然语言问题）。与网页搜索查询不同，企业查询通常存在歧义、因人而异，并且高度依赖上下文，因此系统必须在考虑内容、元数据以及个人和组织上下文的同时，理解用户意图 [4]。例如，在网络搜索中，“Juno 上映日期”这一查询意图明确，指的是2007年的电影 [4], ；而在企业搜索中，同一查询可能有多种意图，例如查找包含名为 Juno 的项目上映日期的文件，或查找来自一位名叫 Juno 的同事的电子邮件。表1还展示了企业搜索查询（即本文的研究重点）与网页搜索查询之间的差异。

因此，企业搜索中的相关性标注本质上比网络搜索更复杂。此外，缺乏公开可用且高质量的企业搜索数据集，无法与现实世界的数据结构和分布相匹配，这限制了企业搜索方法的发展。相比之下，网络搜索得益于大规模的

*两位作者贡献相同。
†通讯作者： {yuekang,jdanovitch,tianweichen,svesal,sosrini}@microsoft.com
‡本研究是在微软任职期间完成的。[4]
https://en.wikipedia.org/wiki/Juno_(film)

公开基准测试集，例如 MS MARCO [5] 和 TREC-CAsT [6], ，推动了检索和相关性标注的快速发展。此外，个性化和上下文化至关重要，因为相关性取决于文本元数据和用户网络。最后，企业查询通常定义不明确、具有领域特异性且属于长尾查询；由于用户参与度低且人工标注数据有限，排序评估尤其具有挑战性。这些挑战凸显了大语言模型在相关性标注方面的价值 [7]。然而，尽管基于大语言模型的判断 [8]已成为企业相关性标注的标准方法，但也存在显著缺点：大语言模型的计算成本高、吞吐量有限，因此不适用于大规模标注任务。为解决这些局限，小语言模型 (SLMs) [9] 提供了一种有前景的替代方案。通过利用精选合成数据进行领域特定微调，小语言模型能够达到与大语言模型相当的不错表现，同时显著提高吞吐量并降低成本。

近期研究已探索训练小语言模型来标注网页搜索查询 [10], ，但企业场景需要更广泛地覆盖结合词汇和语义信号的混合查询，而这一领域仍鲜有研究。为弥补这一空白，我们提出了一种新颖的合成数据生成流水线，通过微调小语言模型，大规模标注企业搜索查询。我们的方法利用企业文档生成真实的用户查询，使用 BM25 [11] 检索高难度负例，并借助基于大语言模型的判断生成分级相关性标签。随后，我们将这些合成数据集提炼到一个小语言模型中，从而得到高效且经济的相关性标注器，能够处理多种企业查询类型。该方法改进了排序模型的离线评估流程，为传统的基于大语言模型的标注器和人工评审员提供了一种可扩展的替代方案。我们的主要贡献如下：

- 我们提出了一种新颖的合成数据生成流水线，利用 BM25 [11] 和基于大语言模型的标注，为企业搜索相关性标注生成高质量、领域专属且真实的训练数据，从而弥补现有数据集的局限。
- 我们展示了对小语言模型 Phi-3.5 MiniInstruct 的有效微调，使其性能达到可与教师大语言模型（GPT-4o）媲美的水平，同时显著提升吞吐量并降低成本。 [12],

这些事实表明 微调后的小语言模型能够以经济高效的方式支持生产规模的离线排序评估 ，提供一种实用的替代方案，可替代基于大语言模型的标注或人工标注。

2 相关工作

大语言模型与小语言模型 生成式语言模型的快速发展改变了自然语言处理（NLP）的面貌，并重塑了现代信息检索（IR），使系统能够在广泛的应用中进行高级推理和理解 [13, 14, 15]。ELMo [16]和 BERT [17] 等早期创新表明，在原始文本上进行大规模预训练能够捕捉丰富的上下文依赖关系；随后，ELECTRA [18] 和 DeBERTa [19] 等改进则引入了更高效、更准确的预训练策略。序列到序列架构的引入，例如 T5 [20], ，实现了对各类信息检索任务的统一处理，涵盖查询改写到段落排序。ColBERT [21],等基于嵌入的检索方法进一步凸显了编码器通过捕捉细微语义信号超越词汇匹配器的能力。与此同时，GPT [22, 23] 等仅解码器模型标志着大规模生成式系统时代的到来，并迅速将应用拓展至文档排序、对话式搜索和摘要生成 [24],，最终确立了其作为现代信息检索最先进引擎的地位。

近年来，关注重点已从更大的模型转向紧凑且性能不俗的小语言模型（SLM）。Gemma [25], Llama [26], 和 Phiseries [9, 12] 等系列表明，参数量仅数十亿的模型也能具备出色的泛化能力和推理能力，同时大幅降低推理成本。尽管小语言模型通常在绝对性能上不及前沿大语言模型，但其速度快、内存占用低、推理成本较低，因此在吞吐量和可扩展性至关重要的检索与标注任务中极具吸引力 [27]。这种以略低的峰值准确率换取效率的权衡，使小语言模型能够成为大语言模型的实用补充，无须耗费大量 GPU 资源，即可支持大规模离线评估和快速实验。

使用语言模型进行相关性标注相关性评估长期以来一直是信息检索的基石，传统上依赖人工标注以及 BM25 [11] 等启发式模型。尽管这些方法准确，但成本高昂且规模有限。近年来，人们开始探索将大语言模型用作替代方案，以复现或增强人工判断。 [28] 提出了用于自动相关性估计的单次标注方法，表明大语言模型能够高效生成标签。 [29] 研究了利用大语言模型填补相关性判断空缺；随后， [8] 表明大语言模型能够准确预测搜索者偏好，并与用户行为高度一致；而 [30] 则提出了基于标准的大语言模型判断，以提高主题相关性和覆盖度等维度上的标签一致性。

查询类型	企业	Web
描述	查询利用上下文和内部实体， 以及语义理解	查询依赖常识和 一般语义
示例	iris 项目日期	虹膜结构
主要区别	进一步需要企业特定的知识- 边缘（例如，“虹膜” =同事/项目）	使用全局/常识知识（例如，“虹膜” =眼睛的一部分）

表1：企业搜索与网络搜索：两者都可以是语义搜索，但企业查询依赖特定领域的知识。

与此同时，LoRA [31] 等微调技术使得将大语言模型适配到领域特定排序任务 [32]成为可能。更高效的基于蒸馏的方法也相继出现： [10] 证明，微调小型大语言模型能够提升增强型相关性数据集的质量，凸显了通过后训练小语言模型实现可扩展评估的潜力。RRADistill [7] 将大型模型蒸馏为小语言模型，用于长尾查询的重排序；Rank1 [33] 则利用基于推理的大语言模型（例如 OpenAI 的 o1 [34], DeepSeek 的 R1 [35]）以及源自 MS MARCO 的轨迹，训练更小、更具可解释性且性能出色的重排序模型。总的来说，这些方法凸显了合成数据流水线与蒸馏在构建紧凑而有效的检索模型方面的潜力。

尽管取得了这些进展，但针对相关性标注的小语言模型探索仍主要局限于开放领域、由语义驱动的检索任务 [10]。这些研究表明，小型模型能够提升数据集质量，或近似语义相关性，但企业搜索面临着独特的挑战。企业环境中的查询通常会将稀疏关键词匹配与语义意图结合起来，而文档集合则涵盖 Teams 聊天、电子邮件和文件等异构领域。目前的工业实践仍依赖前沿级大语言模型，并结合精心设计的提示工程和思维链推理 [36],来应对这种复杂性 [8]。据我们所知，此前尚无研究考察经过微调的小语言模型能否在这一场景中提供具有竞争力且资源高效的标注。我们的工作通过针对企业相关性标注对 Phi-3.5-mini 进行适配，弥补了这一空白，并表明精心整理的训练数据集和蒸馏流程能够使小语言模型成为更快速、更高效且适用于大规模实验的实用替代方案。

利用大语言模型生成合成数据以创建查询，已证明在低资源场景下对模型训练十分有效[37]。仅基于合成数据微调的检索方法，在各种领域特定信息检索任务中均优于 BM25 等强基线以及其他自监督稠密检索方法 [38]。Promptagator[39] 和 InPars v2 [40] 等方法推动了开源排序器的发展，使其在公开基准上取得了最先进的结果；而 DUQGen [41] 则通过聚类和概率采样探索无监督领域自适应，以增加合成查询的多样性。

基于这些创新，我们聚焦于将小语言模型（SLM）标注器实际部署到企业搜索与检索中。具体而言，我们使用通过大语言模型驱动的查询与标注流程生成的领域特定数据对 Phi-3.5 Mini [9] 进行微调，并通过针对合成企业风格查询和文档进行指令微调来定制模型。我们选择 Phi-3.5 Mini 和 GPT-4o 作为基准模型，因为它们在近期的检索与推理研究中表现出色 [24], , 且分别是各自模型类别（紧凑型小语言模型与前沿级大语言模型）的优秀代表。虽然我们未纳入 GPT-5 等较新模型或其他小语言模型系列的结果，但我们认为研究发现具有普遍适用性，因为核心优势源自数据生成与蒸馏流程，而非任何特定模型的属性。最终得到的小语言模型标注器达到了与前沿大语言模型相当的性能，同时显著提升了效率和可扩展性，因此非常适合大规模企业相关性评估。值得注意的是，传统的搜索相关性标注方法（如人工标注或直接依赖前沿大语言模型）面临着重大挑战。由于隐私政策严格，且处理大规模数据集需要付出大量工作，在现实企业搜索场景中，人工标注往往难以实施。基于大语言模型的标注虽然能够生成高质量标签，但计算成本高、速度慢，不适合大规模部署。

2.1 问题定义

给定用户查询 q，目标是确定一组文档 $D = \{d_1, d_2, ..., d_n\}$ 与查询q之间的相关性。这可以形式化为学习一个相关性函数 $r(q, d_i)$，该函数为每个查询-文档对 (q, d_i) 分配一个相关性分数，其中 $d_i \in D$。我们将问题定义如下：

输入：

- 一个查询 $q \in \mathcal{Q}$，其中 $\mathcal{Q}$ 是由所有用户企业查询构成的空间。
- 一组文档 $D = \{d_1, d_2, ..., d_n\}$，其中 $d_i \in \mathcal{D}$ 和 $\mathcal{D}$ 是文档空间。

数据集	# 样本	描述	用途
人工审核企业文档件	1,500	一个包含相关元数据的内部文档专有集合，在人工审核环境下整理而成。	**监督微调：**用于生成合成查询–文档–标签三元组，并据此微调模型。
INTERS [42]	∼250,000 个样本涵盖 20 种不同任务	涵盖多种任务的多任务调优数据集，包括查询、文档以及查询与文档之间关系的理解，。	**监督微调：**用于多任务训-练。（图3中的绿色方框）
TREC-CAsT [6]	4,000	用于对话式信息检索（CIS）的公开数据集包含段落-查询-人工标签三元组。	**监督微调：**使用 ∼4k 个样本进行指令调优。（图3中的蓝色方框）
MS MARCO 段落 [5]	超过 400,000 个段落	用于信息检索的开放数据集，包含段落——查询对，其中一些带有人工标注，标注等级为 0–3。	**监督微调：**使用 GPT-4o 生成用于指令调优的合成查询。（图中的紫色方框图3）
人工标注的企业搜索数据集	923	由专有企业查询-文档对组成的数据集，由经过培训的人工标注员标注相关性。	**评估：**用作黄金标准评估数据集，用于最终比较。

表 2：用于微调和评估的数据集

输出：

- 为每个 $(q,\ d_i)$ 对指定相关性分数 $r(q,\ d_i) \in \mathcal{R}$，其中 $\mathcal{R}$ 是相关性分数空间。本文采用 0–4 的序数尺度，其中 0 表示不相关，4 表示高度相关。分数 $r(q,\ d_i)$ 旨在近似人工判断，并与 GPT-4o 等最先进的大语言模型的质量相符。

我们的目标是开发一种实用且可扩展的方法，用于企业搜索相关性标注，从而在确保真实世界信息过滤系统中监督信号高质量且一致的同时，快速、经济高效地评估各种信息检索模型。

2.2 数据集

表 2 详细描述了这些数据集，并说明了它们在从模型训练到评估的不同阶段中所承担的角色。具体而言，我们使用四个数据集进行训练，并使用一个基准数据集进行最终评估。我们获取了 1,500 份在人工审核环境下整理的专有企业文档，用于合成数据生成和微调；我们认为，任何文档平台都可能获得类似的数据集。由于没有可用于评估的公开企业搜索数据集，我们整理了一个内部基准测试集，其中包含 923 个由经过培训的人工标注员标注的高质量企业查询-文档对。

2.3 评估指标

为了评估我们的小语言模型相关性标注器相较于基线标注器的有效性，我们采用两种互补的评估指标：归一化折损累计增益 (NDCG) [43] 和成对准确率。NDCG 是信息检索领域广泛采用的指标，用于评估排序列表的分级相关性；成对准确率则衡量模型在同一查询下保留文档正确相对排序的能力。这两项指标的值越高，表明模型预测与预期的相关性排序越一致。我们在查询级别计算这两项指标后，报告它们在不同查询上的平均值。

1. 归一化折损累计增益 (NDCG)：我们采用完整的归一化折损累计增益 (NDCG)，以衡量模型在每个查询下对整个排序的表现。NDCG 通过为相关性更高的文档赋予更高增益，并根据其排序位置进行对数折损来衡量排序质量，从而确保将相关条目排在列表靠前位置时获得更高评价。与截断变体（例如 NDCG@k）不同，完整 NDCG 会考虑每个查询检索到的所有文档，从而全面评估排序效果，且这一指标得到广泛认可。

2. 成对准确率指标（准确率）：该指标衡量模型相较于真实标签，能否始终保持文档的正确相对排序。对于每个查询，我们穷举所有可能的文档对（$A,\ B$），并检查模型对它们相关性的相对判断是否与真实情况一致。例如，如果模型和真实标签都表明 A 比 B ($A > B$)更相关，那么这对文档就被计为正确匹配。我们计算一个查询的所有可能组合中正确匹配的文档对所占比例，得出一个介于 0 和 1 之间的分数，然后对所有查询的分数取平均值。表 3 展示了完整的计算矩阵。单元格值为 1 表示模型与真实标签在该比较中完全一致，而 0 则表示两者不一致。

模型输出	$A<B$	$A=B$	$A>B$
模型表示 $A<B$	1	0	0
模型表示 $A=B$	0	1	0
模型表示 $A>B$	0	0	1

表 3：准确率指标 具体信息。此处：1：模型的预测与基线标签完全匹配，0：模型的预测与基线标签不符。

3. 每分钟请求数（RPM）： 我们还报告 RPM，这是一项用于衡量小语言模型标注器在指定硬件上处理请求速率的指标。它通常用于监控系统性能、速率限制和可扩展性。

$$\text{RPM} = \frac{\text{Number of total requests for labeling}}{\text{Time elapsed for the labeling in minutes}}$$

3 方法

与网络搜索不同，网络搜索拥有 MS MARCO [5] 和 TREC-CAsT [6] 等大规模开源数据集，可提供高质量的人工标注查询-文档对；而企业搜索领域却没有公开可用、能够体现其独特特征的数据集。尽管一些企业系统可能会收集查询日志，但并非总是如此，而且这些日志往往包含敏感的机密用户信息，因此无法据此收集人工相关性判断。此外，点击日志等交互数据本质上存在偏差，因为它们主要反映用户参与度，而非真实相关性 [44]。例如，无论结果质量如何，用户往往都会点击排名靠前的结果 [45]。这些局限凸显了对替代方案的需求：既能有效训练，又能解决缺乏可行训练数据的问题。

为此，我们提出两种方法，在不依赖大规模查询日志的情况下构建用于企业相关性标注的微调数据集。这两种方法都只需要一组经过筛选的种子文档。因此，我们的方法完全自动化，且从根本上保护隐私，非常适合用户查询敏感、稀缺或无法获取的企业场景。具体而言，这两种方法都基于这样一种思路：从上述种子文档生成合成查询。为了生成用于微调小语言模型标注器的合成查询，我们研究了两种不同的方法。第一种方法微调一个辅助小语言模型，使其能够根据种子文档和预定义的相关性分数生成查询，从而通过随机采样文档-分数组合来创建合成查询。随后，这些生成的查询将用于训练小语言模型标注器。相比之下，第二种方法使用基于大语言模型的查询生成、基于 BM25 的负例文档挖掘和大语言模型辅助标注，创建合成查询-文档-分数三元组，然后利用这些三元组微调小语言模型标注器。在实践中，我们发现第一种方法难以生成高质量且多样化的查询，尤其不适用于基于关键词的企业搜索，因而实用性有限。因此，我们主要关注第二种方法，它表现出更稳定可靠的效果。下文将详细介绍这两种方法。

3.1 将小语言模型微调用作查询生成器

我们首先将 Phi-3.5 Mini Instruct 模型微调用作查询生成器，因为这种方法能够在控制相关性等级的同时更快地进行大规模查询生成，并减少对昂贵的大语言模型推理的依赖。据我们所知，此前尚无将小语言模型微调用于企业搜索查询生成的研究。此外，我们此前已将同一个小语言模型微调用于语义网查询生成，并以上下文和预定义的相关性分数作为条件。该方法在关键评估指标上取得了令人鼓舞的结果，因此我们决定将其扩展至企业场景。第 1 节介绍了本文重点关注的企业搜索查询与网页搜索查询之间的主要区别。

3.1.1 实验设置

为了生成用于微调 SLM 查询生成器的合成查询，我们利用 GPT-4o，通过精心设计且包含明确多样化示例生成指令的提示词，生成正向和负向企业搜索查询。随后，我们在该合成数据集上采用多种训练配置对 Phi-3.5 Mini Instruct 进行微调，旨在使 SLM 能够根据文档元数据和目标相关性等级生成企业风格查询。然而，当我们通过将微调模型的输出与 GPT-4o 的金标准判断（大语言模型裁判）进行比较来评估该模型时，发现其二元相关性准确率仅为 60.8%；也就是说，其生成的查询中只有 60.8% 与由 GPT-4o 评估的预期相关性类别（相关或不相关）相匹配。进一步的人工分析显示，即使提示词明确要求模型生成不相关（负向）查询，模型仍明显偏向于生成正向或听起来合理的查询，而当提示词

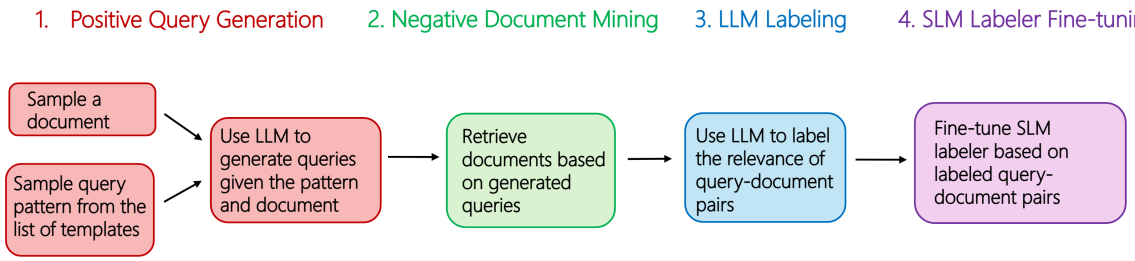

图 1：我们提出的解决方案的整体模型架构。

Example:
File Content: (omitted here due to the space limit)
Metadata: File name: DB Moves budget plan decent.one | Author: John Smith | Parent Folder: Open Projects opportunity | Extension: OneNote | Title: DB Moves budget plan decent
Input pattern: <unk>, 'File Name', 'Author', 'Parent Folder'

Result from 1st LLM call (generation):
Keywords: Cloud, Budget, Module, Security, Container
Queries: Cloud *DB Moves budget* John Open Projects, Module *DB Moves budget* John Open Projects, Security *DB Moves budget* John Open Projects (It uses same segments of different metadata all the time!)

Result from 2nd LLM call (revision for variability):
Revised Queries: Cloud *DB Moves budget* John Open Projects, Module *DB budget plan* John Smith Projects, Security *Moves descent* John Projects

图 2：合成查询生成流程示意图。

明确要求模型生成不相关（负向）查询时。作为示例，为避免隐私问题，我们虚构了一份样本文档，并对 SLM 查询生成器进行评估：

- 内容摘要：一篇介绍如何在 Word 中添加页面的文档。
- 关键词：页面、Word、添加
- 元数据：文件名：“AddPage.docx”，作者：Lisa Morrison，标题：“AddPage”，文件类型：“docx”，父文件夹：“Word 教程”

通过指示模型生成负面查询，我们得到：“Lisa tutorial Docs”、“Add Page”和“Lisa Loop page”，但这三个查询都与该文档相关。
我们认为，这一局限主要归因于两个因素：

- 预训练中的正向偏差：像 Phi-3.5 Mini 这样的指令微调小语言模型经过优化，旨在提供帮助并积极配合。当被要求生成一个“针对不应被检索到的文档的查询”时，即使经过微调，模型仍可能试图生成听起来相关或合理的内容，而非真正无关的内容。
- 模型容量有限：作为一个小模型，Phi-3.5 Mini 可能难以理解细微的指令，例如生成“基于关键词、看似合理但实际无关的查询”；在相关性界限可能较为微妙的企业场景中，这一点尤为明显。

总的来说，尽管微调后的模型能够较好地生成与文档语义信息相符的查询，但它始终难以应对企业搜索查询以关键词为基础的特性。与通常体现用户意图和语义理解的传统网络搜索不同，企业搜索还依赖精确的关键词匹配、文件元数据和结构化文档属性。因此，对小语言模型进行微调无法生成有意义的困难负例和多样化查询，最终也无法很好地应用于真实的企业搜索场景。

3.2 使用大语言模型和检索器进行查询生成

这一替代方法构成了我们工作的核心方法论，将合成数据生成与信息检索技术相结合。具体而言，我们首先使用 GPT-4o 生成并优化合成企业搜索

查询。接下来，我们应用 BM25 等排序算法挖掘合理的负样本文档。基于得到的查询-文档对，我们随后提示 GPT-4o 使用精心设计的标注提示词分配分级相关性标签。最后，我们使用这些查询-文档-标签三元组对 Phi-3.5 Mini Instruct 进行微调。

该合成查询生成流程的整体工作流程如图 1 所示。它包含四个主要组成部分：(1) 正向查询生成，(2) 负样本文档挖掘，(3) 基于大语言模型的标注，以及 (4) 小语言模型标注器微调。下面将详细介绍各个组成部分。

方框 3.1：正向查询生成提示词

Role
Yo你是一名助手，负责根据文件元数据和内容生成**关键词** 和**用户搜索查询** t.

输入
– 文件元数据部分组成的结构化字符串。– 文件的文本内容。– 元数据可能包含内容占位符。

Task
步骤 *1*：关键词提取– 从内容中提取最多 6 个相关的单字关键词。– 排除停用词。– 如果内容为空，则返回空列表。

步骤 *2*：查询生成– 按照元数据字符串中的确切顺序，构造自然的查询。– 将每个内容占位符替换为关键词。– 生成 3 个不同的查询，且不重新排列各部分。

示例
–此处展示少样本示例...
输出格式
关键词：k1、k2、k3、... 查询：
q1、q2、q3

1. 正向查询生成：我们从一个文档数据集和一张查询模式模板表着手。在实践中，可以通过分析用户在企业搜索中通常如何构造查询来建立此类模板。例如，像 <作者姓名><文件名> 这样的模式可能会生成查询“Lisa 预算报告”。

在我们的工作中，我们构建了一张内部模板表，其中包含按排名排列的常见查询模式列表（例如， <文件夹名称><文件类型><关键词>）及其在实际搜索流量中的频率。然而，即使无法访问此类数据，也可以通过系统枚举文档元数据字段与关键词的组合来生成合成模板表；我们发现，这种方法能够产生真实且多样的查询结构。这些查询模式通过对查询文本进行实体解析构建而成，以确保每个片段都对应于文件元数据中的一个实体。此外，模板还可以包含从内容中提取的信息的占位符，将从文本中提取的文档关键词插入其中。因此，它们能够准确反映用户在工作场所环境中通常如何搜索文档。

为确保我们的合成查询能够反映真实的用户行为，我们对查询模式表进行加权抽样，并根据每种模式的观测频率按比例抽取模式（有放回抽样）。因此，频率更高的模式会被更频繁地抽取，从而保留真实企业查询的分布特性。每个抽取出的模式都作为模板使用，系统会动态插入相关的文档元数据（例如作者、标题、文件类型、父文件夹）以及提取出的关键词。

选定一个模式—文档对后，我们利用 GPT-4o 为该组合生成三个不同的企业查询。每个查询生成步骤还包括一个轻量级的关键词摘要阶段，由大语言模型从文档中提取关键主题词，并将其纳入查询中。这有助于使生成的查询立足于文档的实际内容。所使用提示词的结构见方框 3.1。由于保密限制，我们无法披露完整的提示模板。相反，我们提供一个高层次的概要，以说明其总体结构。具体措辞并非我们贡献的核心；相反，我们的关键发现是，微调小型语言模型能够有效地执行企业相关性标注，并适用于不同的提示措辞变体。

对于更复杂的模式，即包含更多组件或罕见元数据组合的模式，我们观察到初始查询有时缺乏变化。例如，如果抽样得到的查询模板很长，大语言模型往往会反复选择每个组件中的同一部分，即使明确要求其进行变化也是如此。为解决这一局限，我们增加了一个基于大语言模型的查询修订阶段，从多样性、流畅性和覆盖度方面对查询进行重写，从而在保留其核心语义相关性的同时，生成一组更丰富、更真实的查询。该查询修订提示词概述见方框 3.2。为说明两个提示词如何依次发挥作用，图 2 展示了一个示例（使用了虚构元数据）。合成查询生成提示会生成三个合理的查询，但它们反复使用相同的元数据片段，从而限制了多样性。随后，精炼步骤根据严格规则重写这些查询，产生更多样化的输出。

此流程生成了高质量的正向合成查询-文档对集合，并使每个文档-模板对平均产生三个多样化且与元数据对齐的查询。

方框 3.2：正向查询修订提示

Role
你是一名负责验证和修订**用户生成的搜索查询**的助手。

输入
– M 元数据字符串（含可能的关键词占位符）– 关键词列表 – 三个生成的查询

Task
步骤 *1*：验证– 确保查询遵循元数据的顺序和结构。 – 将每个内容占位符替换为关键词。 – 避免冗余；让查询简短自然。

步骤 *2*：修改
– 修改查询，确保每个查询中的元数据部分在三个查询中的表述各不相同。

示例
–此处显示少样本示例...

输出格式
修改后的查询：query1、query2、query3

2. 负样本文档挖掘：为构建用于相关性建模的均衡且高质量训练集，我们采用 Okapi BM25 检索算法 [46]，为每个合成查询挖掘负样本文档。BM25 是传统信息检索中广泛使用且表现强劲的基于词汇的基线方法，在缺乏人工点击日志或明确负标签的情况下，它能够帮助我们模拟真实的检索环境。

对于每个合成生成的查询，我们使用 BM25 从语料库中检索排名前-k 个文档，并排除生成该查询所依据的原始源文档。需要特别说明的是，我们并不假设源文档是唯一的相关文档，也不要求所有检索到的文档都严格不相关。相反，我们将检索到的候选文档视为合理的检索结果，它们可能涵盖不同的相关性等级。换言之，在这一阶段，我们的目标并不是预先将检索到的文档分类为相关或不相关，而是收集一个真实的相关性等级范围，并将最终的相关性判断推迟到后续基于大语言模型的标注阶段。

在实践中，BM25 往往会返回与查询在词汇和语义上重合程度各异的文档。某些检索到的文档可能高度相关（例如，讨论同一实体或主题），而另一些则只有部分关联，或基本无关。这种特性非常符合我们的目标：构建涵盖多样化相关性分数的训练数据。因此，检索到的文档自然构成了困难负例、弱正样本和近似未命中案例的混合，反映了现实企业搜索中常见的模糊性。这种多样性对于训练模型区分真正相关的文档与仅在主题或语义上相近的文档至关重要。例如，给定查询“Lisa 预算报告”（该查询源自 Lisa Morrison 撰写的一份文档，其标题为第3.1小节中提到的《2023年预算报告》），BM25 可能会检索到另一份文档，例如《Lisa旅行计划》或《2022年预算草案》。这些结果在词汇上相似，因此看起来是合理的检索结果，但并非预期匹配，因此适合作为训练用的困难负例。

在我们的实验中，我们设置了 $k = 4$，经实证发现，在基于大语言模型的标注后，这能使相关性标签在 0–4 各级别上呈现均匀分布。这种平衡有利于训练稳健的相关性标注器，因为它可以避免

过多地表示显而易见的不相关样例，同时促使模型学习区分相关文档之间的细微差异。在实践中，可根据语料库的规模和多样性动态调整 k，以保持各相关性级别的覆盖度均衡。例如，如果基于大语言模型的标注结果显示较高相关性分数的样例过多，我们可以增大 k，以检索更多候选文档，并仅对新加入的查询-文档对进行标注，同时保留先前已标注的查询-文档对，从而逐步重新平衡数据集，而无需重新处理整个语料库。

3. 大语言模型标注：数据流水线的下一阶段是基于大语言模型的标注，在此阶段，GPT-4o 会为每个查询–文档对分配相关性分数。这一步对于将大语言模型的检索能力提炼到我们的小语言模型标注器中至关重要。我们采用了通用的标注提示模板，指示 GPT-4o 为每个查询–文档对输出离散的相关性分数，范围从 0（不相关）到 4（高度相关），如框 3.3 所示。

为确保准确率和一致性，我们在初始标注后增加了一个质量控制轮次。在此阶段，我们自动检查 GPT-4o 的输出是否存在格式问题和语义不一致。在大约 24,000 个样本中，只有 15 个案例包含无效或不完整的输出（例如缺少分数），这些案例均被自动重新标注。此外，我们还实施了标注后过滤机制，以保持数据集质量。具体而言，如果第 1 步生成的合成正向查询获得了异常低的分数（0 或 1），我们会使用相同的提示词通过 GPT-4o 对其重新标注。如果低分仍然存在，我们将其保留为有效的负例；否则，则丢弃该查询。这一过滤流程可以防止模型学习到相互矛盾的信号，并确保训练分布的一致性。

完成这 四个步骤后，我们获得了一组高质量的合成查询-文档-标签三元组， h
它们构成了 用于微调我们的小语言模型标注器的核心训练数据集。

4. 小语言模型标注器微调：最后，我们使用已标注的查询-文档对对 Phi-3.5 Mini Instruct 模型进行了微调。为了优化指令微调过程，我们沿用了上一步用于大语言模型标注的相同提示词模板，但删除了要求模型在生成预测相关性分数的同时给出解释的指令。通过实证分析，我们发现，尽管解释有助于提高可解释性，但将其纳入会导致训练信号不一致，并使微调后小语言模型的输出变得不必要地冗长。为了进一步提升微调后小语言模型的泛化能力，我们采用了多任务调优 [47]。如图3所示，此阶段加入了若干额外数据集，使模型接触更广泛的查询-文档相关性场景。具体而言，我们纳入了：

- INTERS [42]：该数据集包含针对查询和文档理解的 19 项不同小任务的数据集。我们使用它进行多任务调优，以增强模型的泛化能力。

- TREC-CAsT [6]：该数据集包含语义网查询-段落对，并由人工按 0–4 的等级进行标注。

- MS MARCO段落 [5]：该数据集包含来自各种来源的 40 多万个段落。我们基于思维链思路，从随机选取的段落出发，使用 GPT-4o 生成合成查询，并按 0–4 的相关性等级进行标注。

尽管我们的主要目标是为企业查询标注，我们仍有意纳入了来自公共领域的语义网查询数据集。我们认为，这种更广泛的监督信号有助于模型更深入地理解查询与文档之间的相关性，而不局限于严格的关键词匹配场景，从而提升泛化能力和鲁棒性。具体而言，如图3所示，我们首先使用 INTERS 对 Phi-3.5 Mini Instruct 进行多任务调优，以增强其鲁棒性和整体能力；随后，又在我们上文提到的三个数据集的组合上持续微调该模型：合成的企业查询-文档-标签三元组、带人工标注的 TREC-CAsT，以及包含合成查询的 MS MARCO段落。

我们使用最大序列长度 4096 对模型进行 2 个 epoch 的训练。通过将每台设备的批次大小设为 4，并使用 8 个梯度累积步骤，有效批次大小达到 32。每 40 步记录一次训练日志。每 80 步使用单独的人工标注测试数据集（评估集）进行评估。该评估集不会影响训练过程，因为没有早停机制，也没有来自评估集的反馈循环。训练和评估时，每台设备的批次大小均为 4。微调在由八块 A100 GPU 组成的集群上进行。在该 GPU 集群上，训练速度约为每秒 5.84 个样本。

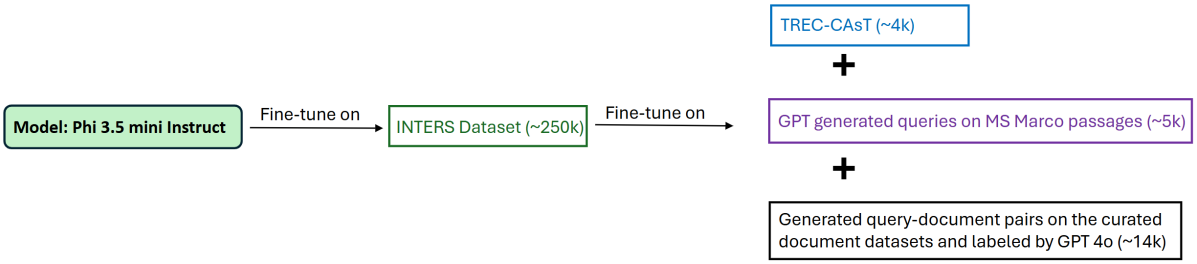

图3：小语言模型标注器监督微调（SFT）工作流程：我们首先在 INTERS 数据集上对 Phi-3.5 Mini Instruct 进行多任务预训练微调，然后再结合合成数据集和开源数据集，对得到的模型进行微调。

框3.3：相关性标注提示

Role
你是一名企业搜索质量评估员，负责评估文件/消息相关性。

Task
给定一个**查询**和一个**实体**（文档），请在 0–4 的范围内为相关性评分：– 4 = 理想质量，应为理想结果..... – 0 = 质量差，绝不应展示

输入
– 查询文本 – 文档元数据和重点内容

输出
得分：<介于 0 和 4>之间

4 实验结果

本节将详细评估微调后的 Phi-3.5 Mini Instruct 模型的性能，并将其与基础模型（原始 Phi-3.5 Mini Instruct）和 GPT-4o 进行比较。结果表明，微调显著提升了 Phi-3.5 Mini 模型与人工相关性判断的一致性，在两项评估指标上均达到甚至略微超过 GPT-4o。

微调的影响：微调对模型性能产生了显著影响。如图 4 所示，SLM-Human NDCG 从 0.815 提升至 0.953，表明模型与人工判断的一致性显著提高。同样，成对准确率从 42.15 提升至 63.81，在两项关键指标上甚至超过了 GPT-4o（NDCG：0.944，准确率：62.58）。这证实了将 Phi-3.5 Mini 适配于企业相关性任务不仅能取得具有竞争力的表现，甚至能实现更优性能，也凸显了合成数据生成和微调流程的有效性。

我们还针对训练数据进行了全面的消融研究，结果汇总于表4。我们的主要发现如下：

数据规模的影响：人们普遍认为，更多的训练数据可能带来更好的结果；然而，我们的实验表明，这种效果在达到一定规模后便趋于平缓。微调结果凸显了数据规模对模型性能的影响。如表4所示，将合成训练数据从 14K 个样例增加到 24K 个样例（第2行与第3行相比），带来的提升微乎其微。使用 14K 个样例进行微调后，SLM-Human NDCG 显著提升（从 0.815 升至 0.953），成对准确率也大幅提高（从 42.15 升至 63.81）。然而，将数据扩展到 24K 个样例后，性能几乎没有变化（NDCG：0.954，准确率：63.55）。比较第4行和第5行（两者均未进行多任务调优）时，也能观察到类似趋势。这表明，尽管扩大数据规模起初能带来显著提升，但超过某个阈值后，继续增加合成数据的收益会逐渐递减。

多任务调优：多任务调优显著提升了模型输出与人工相关性判断的一致性。如表4所示，对比第2行（采用多任务调优）和第4行（未采用），可以清楚地看到多任务调优的影响。在使用相同的 14K 合成训练数据集时，采用多任务调优的模型 NDCG 达到 0.953，准确率达到 63.81；未采用多任务调优时，这两项指标分别为 0.946 和 62.41。这一提升表明，多任务调优结合指令调优，能够有效增强模型对人工相关性偏好的泛化能力。在 24K 合成数据集下对比第3行（采用多任务调优）和第5行（未采用）时，也呈现出类似趋势：NDCG 从 0.945 提高到 0.954，准确率从 62.70 提高到 63.55。

为了进一步验证合成查询的作用，我们在微调模型时不纳入合成企业查询数据集，同时保持其他所有开源数据集不变（即排除图3中黑框内的数据）。如表4第7行所示，该变体相比原始模型仅有小幅变化：归一化折损累计增益从0.815升至0.826，准确率从42.16升至42.88。这一结果符合我们的预期：现有开源数据集主要针对网页搜索查询理解而设计，因此并不能完全体现企业搜索相关性标注的特征。这些结果证实，我们的合成企业查询提供了主要训练信号，而多任务调优则通过提升鲁棒性带来辅助收益。

训练数据集质量：如第3.1小节所述，我们通过查询精炼步骤对大语言模型生成的查询进行精炼，以提升其可信度和多样性。在合成查询生成后应用这一精炼步骤时，模型（表4第2行）表现出色，归一化折损累计增益为0.953，准确率为63.81。相比之下，直接复用相同的1.4万条原始合成查询而不进行精炼（第6行），则导致性能出现明显下降（归一化折损累计增益为0.943，准确率为60.97）。这凸显了训练数据集质量的关键重要性。此外，将数据集规模从1.4万条增至2.4万条样例（第2行与第3行对比），结果几乎相同。这表明，对于小语言模型的微调，一旦数据集超过合理规模的阈值，提升数据质量比单纯增加数据量更为有效。

吞吐量分析： 我们的微调小语言模型在单张 A100 GPU 上达到每分钟 873.33 个请求，推算在由 8 张 A100 GPU 组成的集群上可达到近每分钟 7,000 个请求。这一吞吐量比典型的大语言模型标注器快一个数量级；后者通常仅能达到每分钟数百个请求。

成本分析： 成本是一个重要考量因素，因为大规模评估和训练数据标注都十分频繁且成本高昂。与 GPT-4o 等大语言模型相比，我们微调后的 Phi-3.5 Mini Instruct 具有显著优势，其输入词元和输出词元的成本都便宜近 19× 。具体而言，每 100 万个输入词元的成本为 0.13 美元，而 GPT-4o 为 2.50 美元；每 100 万个输出词元的成本为 0.52 美元，而 GPT-4o 为 10.00 美元 [48]。这种成本效益，加上与人工判断的高度一致性，使经过训练的小语言模型成为可扩展企业相关性标注的理想选择。

统计显著性分析 为进一步验证我们微调后的标注器的有效性，我们针对 GPT-4o 基线进行了统计显著性检验。我们的基准数据集包含 923 个带有人工标注的查询-文档对，涵盖用于评估的 228 个不同查询。对于每个查询，我们分别计算了微调标注器和 GPT-4o 相对于人工标注的成对准确率和归一化折损累计增益。设 $\delta_q =$ 标注器$_q$ −GPT-4o$_q$ 表示每个查询上的性能差异。

由于我们的目标是证明，与 GPT-4o 标注器相比，我们的标注器在吞吐量和成本显著更低的情况下仍能达到相当的性能，因此我们使用 Wilcoxon符号秩检验进行了单侧配对非劣效性检验 [49]。零假设认为，与基线相比，我们的模型表现差距超过预先设定的非劣效性界值 Δ，即，

$$H_0 : \mathbb{E}(\delta_q) \leq -\Delta, \qquad H_1 : \mathbb{E}(\delta_q) > -\Delta.$$

我们采用较为保守的非劣效性界值：成对准确率为 $\Delta = 0.1\%$，归一化折损累计增益为 $\Delta = 0.0001$ ，相对于各自指标的取值范围，这些界值所代表的差异可以忽略不计。

单侧 Wilcoxon符号秩检验所得的 p值分别为：成对准确率 0.012，归一化折损累计增益 0.00098，二者均低于 0.05 的显著性水平。因此，我们拒绝这两个指标对应的零假设，这表明在设定的较小界值内，我们的微调标注器不劣于 GPT-4o 基线。这些成对准确率和归一化折损累计增益结果提供了统计证据，表明我们的微调模型达到了与强大的 GPT-4o 标注器相当的性能。

定性分析： 检查人工标注的评估集中的示例后，我们发现，经过微调的标注器相比原始小语言模型有了明显提升。具体而言，原始小语言模型甚至 GPT-4o 往往会避免给出最低的相关性分数，即使文档完全偏离主题也是如此，从而导致相关性判断过于乐观。相比之下，我们经过微调的查询标注器校准能力更强，也更有信心：它

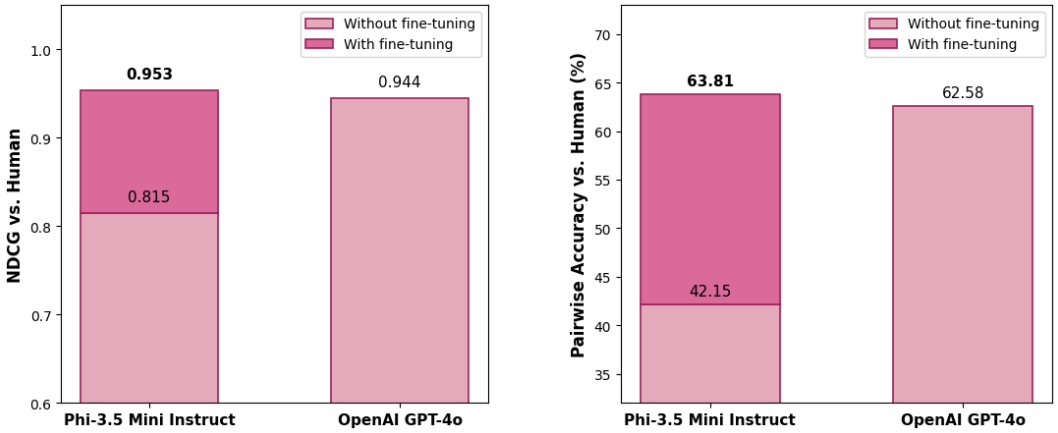

图 4：原始小语言模型和微调小语言模型与 GPT-4o 在企业搜索测试数据集上的两项关键评估指标上的性能结果对比。

Row	多任务调优	数据修订	合成数据规模	小语言模型-人工		小语言模型-大语言模型		大语言模型-人工	
				NDCG	准确率（%）	NDCG	准确率（%）	NDCG	准确率（%）
1	–	–	–（原始小型语言模型）	0.815	42.15	0.820	41.38		
2	✓	✓	14K	0.953	63.81	0.951	66.16		
3	✓	✓	24K	0.954	63.55	0.950	66.10	0.944	62.58
4	✗	✓	14K	0.946	62.41	0.950	66.72		
5	✗	✓	24K	0.943	62.50	0.948	66.83		
6	✗	✗	14K	0.943	60.97	0.948	66.05		
7	✓	✓	0（仅使用公开数据）	0.826	42.88	0.824	42.19		

表4：比较不同微调配置下原始小语言模型与训练后小语言模型性能的结果，用于消融研究。最后三组列展示了基于小语言模型/小语言模型/大语言模型与人工/大语言模型/人工标注的成对指标。

能够在段落完全不相关时正确赋予 0 分，从而更贴近人工标注。

这些结果共同表明，使用高质量合成数据对小语言模型进行精心微调，不仅能使其与人工判断的对齐程度达到可与大型大语言模型媲美的水平，还能显著提升相关性标注的可扩展性和成本效益。

5 结论

我们的工作通过使用合成数据生成流水线微调小语言模型（SLM），提出了一种高效且可扩展的企业搜索相关性标注方法。经过微调的 Phi-3.5 Mini Instruct 达到了 GPT-4o 级别的准确率，同时显著提高了吞吐量并降低了成本，使大规模查询-文档相关性标注切实可行。我们利用 GPT-4o 生成查询、使用 BM25 进行困难负样本挖掘，并采用基于大语言模型的判断进行标注，从而构建高质量的合成训练数据，使小语言模型能够与大语言模型一较高下。

实证结果表明，我们的微调模型在小语言模型人工归一化折损累计增益（NDCG）指标上达到 0.953，成对准确率达到 63.81，优于 GPT-4o（NDCG：0.944，准确率：62.58）。这些发现凸显了经过良好训练的紧凑型小语言模型能够在与人工相关性判断对齐方面媲美甚至超越大型模型，同时实现更快速、更可持续的标注流水线。更广泛地说，这项工作彰显了合成数据驱动训练弥合高质量、与人工判断对齐的评估和可扩展模型开发之间差距的潜力，为在数据密集型自然语言处理任务中部署轻量级模型提供了一条切实可行的路径。

参考文献

[1] Yutao Zhu、Huaying Yuan、Shuting Wang、Jiongnan Liu、Wenhan Liu、Chenlong Deng、Haonan Chen、Zheng Liu、Zhicheng Dou 和 Ji-Rong Wen。《大语言模型在信息检索中的应用：综述》，2024年。

[2] Cunxiang Wang、Xiaoze Liu、Yuanhao Yue、Xiangru Tang、Tianhang Zhang、Cheng Jiayang、Yunzhi Yao、Wenyang Gao、Xuming Hu、Zehan Qi、Yidong Wang、Linyi Yang、Jindong Wang、Xing Xie、Zheng Zhang 和 Yue Zhang。《大语言模型事实性的综述：知识、检索与领域特异性》，2023年。[3] David Hawking。《企业搜索面临的挑战》。载于*ADC*，第4卷，第15–24页，2004年。[4] Filip Radlinski 和 Nick Craswell。《对话式搜索的理论框架》。载于*2017*年人类信息交互与检索会议论文集，第117–126页，2017年。[5] Tri Nguyen、Mir Rosenberg、Xia Song、Jianfeng Gao、Saurabh Tiwary、Rangan Majumder 和 Li Deng。《MS MARCO：人工生成的机器阅读理解数据集》，2016年。[6] Jeffrey Dalton、Chenyan Xiong 和 Jamie Callan。《TREC CAST 2019：对话式辅助赛道概述》。*arXiv*预印本 *arXiv:2003.13624*，2020年。[7] Nayoung Choi、Youngjune Lee、Gyu-Hwung Cho、Haeyu Jeong、Jungmin Kong、Saehun Kim、Keunchan Park、Sarah Cho、Inchang Jeong、Gyohee Nam 等。《RRADistill：提炼大语言模型的段落排序能力，用于搜索引擎中的长尾查询文档重排序》。载于*2024*年自然语言处理实证方法会议：产业赛道论文集，第627–641页，2024年。[8] Paul Thomas、Seth Spielman、Nick Craswell 和 Bhaskar Mitra。《大语言模型能够准确预测搜索者偏好》。载于第*47*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，第1930–1940页，2024年。[9] Marah Abdin、Jyoti Aneja、Hany Awadalla、Ahmed Awadallah、Ammar Ahmad Awan 等。《Phi-3技术报告：可在手机本地运行的高性能语言模型》，2024年。[10] Quentin Fitte-Rey、Matyas Amrouche 和 Romain Deveaud。《利用微调小型大语言模型构建增强型相关性数据集》。*arXiv*预印本 *arXiv:2504.09816*，2025年。[11] Stephen E Robertson、Steve Walker、Susan Jones、Micheline M Hancock-Beaulieu、Mike Gatford 等。*Okapi* 在 *TREC-3* 中的应用。英国图书馆研究与开发部，1995年。[12] Marah Abdin、Jyoti Aneja、Hany Awadalla、Ahmed Awadallah、Ammar Ahmad Awan、Nguyen Bach、Amit Bahree、Arash Bakhtiari、Jianmin Bao、Harkirat Behl 等。《Phi-3技术报告：可在手机本地运行的高性能语言模型》。*arXiv*预印本 *arXiv:2404.14219*，2024年。[13] Humza Naveed、Asad Ullah Khan、Shi Qiu、Muhammad Saqib、Saeed Anwar、Muhammad Usman、Naveed Akhtar、Nick Barnes 和 Ajmal Mian。《大语言模型综合概述》。*ACM*智能系统与技术汇刊，16(5):1–72，2025年。[14] Shijie Wu、Ozan Irsoy、Steven Lu、Vadim Dabravolski、Mark Dredze、Sebastian Gehrmann、Prabhanjan Kambadur、David Rosenberg 和 Gideon Mann。《BloombergGPT：面向金融的大语言模型》，2023年。网址 *https://arxiv. org/abs/2303.17564*，2024年。[15] Arun James Thirunavukarasu、Darren Shu Jeng Ting、Kabilan Elangovan、Laura Gutierrez、Ting Fang Tan 和 Daniel Shu Wei Ting。《医学领域中的大语言模型》。《自然·医学》，29(8):1930–1940，2023年。[16] Matthew E. Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee 和 Luke Zettlemoyer。《深度上下文化词表示》。载于 Marilyn Walker、Heng Ji 和 Amanda Stent 编，*2018*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇论文），第2227–2237页，美国路易斯安那州新奥尔良，2018年6月。计算语言学协会。[17] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。《BERT：用于语言理解的深度双向变换器预训练》。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，2019年。[18] Kevin Clark、Minh-Thang Luong、Quoc V. Le 和 Christopher D. Manning。《ELECTRA：将文本编码器作为判别器而非生成器进行预训练》。载于*ICLR*，2020年。[19] Pengcheng He、Xiaodong Liu、Jianfeng Gao 和 Weizhu Chen。《DeBERTa：采用解耦注意力的解码增强型 BERT》。载于国际学习表征会议，2021年。[20] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。《探索统一文本到文本变换器迁移学习的极限》。机器学习研究期刊，21(140):1–67，2020年。[21] Omar Khattab 和 Matei Zaharia。《ColBERT：通过 BERT 上的上下文化晚期交互实现高效且有效的段落搜索》。载于第*43*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，第39–48页，2020年。

[22] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Nee- lakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。《语言模型是少样本学习器》。神经信息处理系统进展，33:1877–1901，2020年。[23] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。《GPT-4o 系统卡》。*arXiv* 预印本 *arXiv:2410.21276*，2024年。[24] Yutao Zhu、Huaying Yuan、Shuting Wang、Jiongnan Liu、Wenhan Liu、Chenlong Deng、Haonan Chen、Zheng Liu、Zhicheng Dou 和 Ji-Rong Wen。《用于信息检索的大语言模型：综述》。*arXiv* 预印本 *arXiv:2308.07107*，2023年。[25] Gemma 团队、Morgane Riviere、Shreya Pathak、Pier Giuseppe Sessa、Cassidy Hardin、Surya Bhupatiraju、L´eonard Hussenot、Thomas Mesnard、Bobak Shahriari、Alexandre Ram´e 等。《Gemma 2：以实用规模改进开放语言模型》。*arXiv* 预印本 *arXiv:2408.00118*，2024年。[26] Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等。《Llama 3 模型家族》。*arXiv* 预印本 *arXiv:2407.21783*，2024年。[27] Chris Samarinas 和 Hamed Zamani。《面向文档重排序的小语言模型推理蒸馏与精炼》。载于*2025*年国际 *ACM SIGIR* 信息检索创新概念与理论会议（*ICTIR*）论文集，第430–435页，2025年。[28] Sean MacAvaney 和 Luca Soldaini。《用于自动相关性估计的单次标注》。载于第*46*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，第2230–2235页，2023年。[29] Zahra Abbasiantaeb、Chuan Meng、Leif Azzopardi 和 Mohammad Aliannejadi。《我们能否使用大语言模型填补相关性判断空缺？》*arXiv* 预印本 *arXiv:2405.05600*，2024年。[30] Naghmeh Farzi 和 Laura Dietz。《基于标准的大语言模型相关性判断》。载于*2025*年国际 *ACM SIGIR* 信息检索创新概念与理论会议（*ICTIR*）论文集，第254–263页，2025年。[31] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。《LoRA：大语言模型的低秩适配》。*ICLR*，1(2):3，2022年。[32] Navid Mehrdad、Hrushikesh Mohapatra、Mossaab Bagdouri、Prijith Chandran、Alessandro Magnani、Xunfan Cai、Ajit Puthenputhussery、Sachin Yadav、Tony Lee、ChengXiang Zhai 等。《用于商品搜索相关性判断的大语言模型》。*arXiv* 预印本 *arXiv:2406.00247*，2024年。[33] Orion Weller、Kathryn Ricci、Eugene Yang、Andrew Yates、Dawn Lawrie 和 Benjamin Van Durme。《Rank1：信息检索重排序的测试时计算》。*arXiv* 预印本 *arXiv:2502.18418*，2025年。[34] Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、Ahmed El-Kishky、Aiden Low、Alec Helyar、Aleksander Madry、Alex Beutel、Alex Carney 等。《OpenAI o1 系统卡》。*arXiv* 预印本 *arXiv:2412.16720*，2024年。[35] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。《DeepSeek-R1：通过强化学习激发大语言模型的推理能力》。*arXiv* 预印本 *arXiv:2501.12948*，2025年。[36] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。《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，35:24824–24837，2022年。[37] Marco Braga、Pranav Kasela、Alessandro Raganato 和 Gabriella Pasi。《利用大语言模型为个性化社区问答生成合成数据》。载于*2024*年 *IEEE/WIC* 国际网络智能与智能体技术会议（*WI-IAT*），第360–366页，2024年。[38] Luiz Bonifacio、Hugo Abonizio、Marzieh Fadaee 和 Rodrigo Nogueira。《InPars：面向信息检索的无监督数据集生成》。载于第*45*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，SIGIR '22，第2387–2392页，美国纽约州纽约市，2022年。计算机协会。[39] Zhuyun Dai、Vincent Y Zhao、Ji Ma、Yi Luan、Jianmo Ni、Jing Lu、Anton Bakalov、Kelvin Guu、Keith Hall 和 Ming-Wei Chang。《Promptagator：基于8个示例的少样本稠密检索》。载于第十一届国际学习表征会议，2023年。[40] Vitor Jeronymo、Luiz Bonifacio、Hugo Abonizio、Marzieh Fadaee、Roberto Lotufo、Jakub Zavrel 和 Rodrigo Nogueira。《InPars-v2：将大语言模型用作高效的信息检索数据集生成器》，2023年。

[41] Ramraj Chandradevan、Kaustubh Dhole 和 Eugene Agichtein。《DUQGen：通过多样化合成查询生成，有效实现神经排序器的无监督领域自适应》。载于 Kevin Duh、Helena Gomez 和 Steven Bethard 编，《style id='2'>2024年北美计算语言学协会分会会议论文集：人类语言技术（第1卷：长篇论文），第7430–7444页，墨西哥城，墨西哥，2024年6月。计算语言学协会。[42] Yutao Zhu、Peitian Zhang、Chenghao Zhang、Yifei Chen、Binyu Xie、Zheng Liu、Ji-Rong Wen 和 Zhicheng Dou。《INTERS：通过指令微调，释放大语言模型在搜索中的潜力》。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第2782–2809页，曼谷，泰国，2024年8月。计算语言学协会。[43] Yining Wang、Liwei Wang、Yuanzhi Li、Di He 和 Tie-Yan Liu。《归一化折损累计增益类排序指标的理论分析》。载于学习理论会议，第25–54页。PMLR，2013年。[44] Ali Vardasbi、Harrie Oosterhuis 和 Maarten de Rijke。《逆倾向评分失效时：无偏学习排序的仿射校正》。载于第*29*届*ACM*信息与知识管理国际会议论文集，第1475–1484页，2020年。[45] Nick Craswell、Onno Zoeter、Michael Taylor 和 Bill Ramsey。《点击位置偏差模型的实验比较》。载于*2008*年国际网络搜索与数据挖掘会议论文集，第87–94页，2008年。[46] Stephen Robertson、Hugo Zaragoza 等。《概率相关性框架：BM25及其他方法》。信息检索基础与趋势®，3(4):333–389，2009年。[47] Shengyu Zhang、Linfeng Dong、Xiaoya Li、Sen Zhang、Xiaofei Sun、Shuhe Wang、Jiwei Li、Runyi Hu、Tianwei Zhang、Fei Wu 等。《大语言模型的指令微调：综述》。*arXiv* 预印本 *arXiv:2308.10792*，2023年。[48] OpenAI API 定价。https://openai.com/api/pricing/，2025年。[49] Frank Wilcoxon。《基于排序方法的个体比较》。载于《统计学突破：方法论与分布》，第196–202页。Springer，1992年。