重排序器作为相关性评判者

Chuan Meng
爱丁堡大学 爱丁堡，英国
chuan.meng@ed.ac.uk

Jiqun Liu
俄克拉荷马大学 诺曼，美国 jiqunliu@ou.edu

Mohammad Aliannejadi
阿姆斯特丹大学 阿姆斯特丹，荷兰
m.aliannejadi@uva.nl

Fengran Mo
蒙特利尔大学 蒙特利尔，加拿大
fengran.mo@umontreal.ca

Jeff Dalton
爱丁堡大学 爱丁堡，英国
jeff.dalton@ed.ac.uk

Maarten de Rijke
阿姆斯特丹大学，阿姆斯特丹，荷兰 m.derijke@uva.nl

arXiv:2601.04455v1 [cs.IR] 8 Jan 2026

摘要

使用大语言模型（LLM）预测相关性判断已展现出良好的效果。大多数研究将这项任务视为一条独立的研究方向，例如，侧重于提示设计，以根据查询和段落预测相关性标签。然而，预测相关性判断本质上属于相关性预测，这类问题已在重排序等任务中得到广泛研究。尽管两者可能存在重叠，却鲜有研究探索复用或适配成熟的重排序方法来预测相关性判断，可能导致资源浪费和重复开发。为弥合这一空白，我们在重排序器作为相关性评判器 的设置下复现了重排序器。我们设计了两种适配策略：（i）将重排序器生成的二元标记（如“true”和“false”）直接用作判断；（ii）通过阈值化将连续重排序分数转换为二元标签。我们使用来自3个家族、规模从2.2亿到320亿参数不等的8种重排序器，在TREC-DL 2019–2023上进行了广泛实验，并分析了基于重排序器的评判器所表现出的评估偏差。结果表明，在这两种策略下，基于重排序器的相关性评判器在约40%–50%的情况下优于最先进的基于大语言模型的相关性评判器 UMBRELA；它们还表现出明显的自我偏好，倾向于自身及同一家族的重排序器，同时也存在跨家族偏差。

CCS 概念

• **信息系统** →**检索结果评测**。

关键词

重排序、相关性预测、相关性判断

1 引言

相关性判断将每个查询映射到应为其检索出的文档 [49], , 在信息检索（IR）中发挥着关键作用。准确的相关性判断对于训练和评测排序系统至关重要 [48]。然而，人工标注相关性判断需要耗费大量人力 [3]。近年来，信息检索社区见证了大语言模型（LLM）自动预测相关性判断的应用激增 [4, 16, 19, 21, 33, 47, 52, 56],, 并已展现出令人鼓舞的效果 [52]。

大多数研究倾向于将基于大语言模型的相关性判断视为一个独立的研究方向 [19]。然而，Soboroff [49] 认为，“让计算机判断一篇文档是否相关，与使用计算机检索文档并按照预测的相关程度对其进行排名并无不同”，而这两者“本质上都是对相关性的预测”。同样，Clarke 和 Dietz [9]认为，“基于大语言模型的相关性评估与基于大语言模型的重排序方法之间没有根本区别。”

研究动机。本研究的出发点有两个关键动机：(i) 尽管基于大语言模型的相关性判断预测与排序在概念上存在重叠，但鲜有研究探讨如何将成熟的排序方法（例如重排序器 [27, 38], ）改造用于预测相关性判断。若在构建基于大语言模型的评估器时忽视这些排序方法，可能会造成重复工作，并导致成熟的相关性预测资源未得到充分利用。(ii) 尽管许多研究 [9, 17–19, 49]警告称，同时使用同一个大语言模型进行排名和评测可能导致“循环性”（即排名性能被过度夸大），但目前缺乏对此问题进行探究的实证研究。

研究目标。本文考察关于重排序的既有发现，在多大程度上能够推广到“重排序器作为相关性评估器”这一设置。换言之，本研究在一种新设置下（即作为排名评估器）复现现有的重排序方法。我们重新审视重排序文献中的以下发现：(i) 基于较大语言模型的重排序器通常优于基于较小模型的重排序器 [27, 38, 57]；(ii) 基于推理的重排序器会在作出相关性预测之前生成明确的推理链，其表现通常优于不基于推理的重排序器 [57, 58]；(iii) 基于推理的重排序器表现出较强的稳健性 [58]。

挑战与解决方案。主要挑战在于相关性判断预测与重排序之间存在输出格式不匹配：前者为查询-文档对分配离散标签（例如，相关或不相关），而后者通常输出连续相关性分数，以生成排序列表。为弥合这一差距，我们设计了两种适配策略：(i) 直接生成：对于根据特殊标记（例如“true”和“false”）的输出逻辑值计算相关性分数的重排序器 [38, 57], , 我们直接使用最终生成的标记作为预测相关性判断；以及 (ii) 分数阈值化：通过阈值将连续重排序分数转换为二元标签。

范围。基于上述介绍的两种适配策略，本研究考察每种策略下基于重排序器的相关性评判器的有效性。我们首先探讨以下研究问题：
RQ1： 重排序器能否通过直接生成泛化为有效的相关性评判

本作品采用知识共享署名 4.0 国际许可协议授权。

器？ **RQ2：** 重排序器采用分数阈值化时表现如何？除有效性之外，将重排序器用作相关性评判器还可能引入评估偏差和循环性 [9, 17–19, 49]。因此，有必要研究基于重排序器的评审器所引入的偏差风险，我们进一步考察以下研究问题： **研究问题 3：** 基于重排序器的评审器在评估重排序器时，在多大程度上表现出偏差？

实验。我们在 TREC 2019–2023 深度学习（TREC-DL）赛道 [10–14], 上开展实验，这些赛道提供了密集且高质量的人工标注相关性判断。我们的研究考察了 3 个具有代表性的重排序器系列：monoT5 [38],Rank、LLaMA [27], ，二者在排序研究文献中均得到广泛使用 [30, 57, 58],；以及 Rank1 [57], ，一种近期提出的、达到先进水平的推理型重排序器。我们总共纳入了这 3 个系列中的 8 个重排序器，模型规模从 2.2 亿到 320 亿个参数不等。我们通过测量这些基于重排序器的相关性评判器与 TREC 评审员之间的一致性（使用 Cohen's κ），并比较系统排名，来评估它们。为了给基于重排序器的相关性评判器提供参照，我们使用了 UMBRELA [55, 56], ，一种先进的基于大语言模型的相关性评判器。

启示。对于 RQ1 和 RQ2，我们发现，在这两种适配策略下，大多数关于重排序的发现都适用于“重排序器作为相关性评判器”的设置。我们发现，在约 40%–50% 的案例中，基于重排序器的相关性评判器的表现优于 UMBRELA [55, 56], 。基于推理的重排序器（Rank1）通常优于非推理型重排序器，并且在不同数据集上的稳健性更强。扩大模型规模往往有助于提升系统排名质量，但这种提升并不总是一致的。分数阈值化对相关性分数高度集中于两端的重排序器尤为有效；与其他重排序器相比，Rank1 对阈值变化表现出显著更强的稳健性。对于研究问题 3，我们发现，基于重排序器的评审器存在明显偏差：它们对自身重排序器的评分最高，其次是同一家族的重排序器，最后是其他重排序器；它们表现出跨家族偏差（例如，非推理型重排序器更偏好其他非推理型家族，而非推理型Rank1系列），并且与 UMBRELA 相比，往往会高估 BM25 的检索质量。

总的来说，基于重排序器的评审器所展现出的竞争力，为相关性判断并非一项全新的任务，而是相关性预测的一种特定情形（与 [9, 49]一致）提供了经验证据。这表明，社区可以基于成熟的相关性预测器构建相关性评判器，以减少重复工作。同期研究 [23] 通过添加可训练参数，将 monoT5 适配为评审器。

贡献。我们的主要贡献如下：

- 据我们所知，我们是首个系统地复现将重排序器用作相关性判断预测器的研究。
- 我们设计了两种适配策略——直接生成和分数阈值化，使重排序器能够充当相关性评判器。
- 实验结果表明，基于重排序器的相关性评判器在约 40%–50% 的情况下，性能优于 UMBRELA，但往往会对自身及同系列重排序器表现出强烈的自我偏好，也存在跨系列偏差。我们已在以下地址开源补充材料：https: // github.com/ChuanMeng/reranker-as-judge。

2 任务定义：从重排序到相关性判断

鉴于逐点重排序器在信息检索领域已得到充分研究, [27, 38, 57], 且许多广泛使用的信息检索评估指标（如精确率、召回率、平均倒数排名和平均精度均值）通常用于二元相关性判断，因此本文聚焦于逐点重排序器和二元相关性判断。

给定一个查询 q 和一个文档 d，重排序器 f 会赋予其连续的相关性分数 $s = f(q, d) \in \mathbb{R}$，并根据这些分数对文档进行排序，从而生成最终排名。为了将重排序器 f 适配为相关性评判器， f 则会输出离散的相关性标签 $l = f(q, d) \in \{0, 1\}$，其中 $l = 1$ 表示 d 与 q 相关，否则则表示不相关。

3 可复现性方法

3.1 研究问题与实验设计

RQ1 重排序器通过直接生成能否泛化为有效的相关性评判器?

为回答 **RQ1**，我们采用所提出的直接生成策略（详见下文第 3.2.2 节），将重排序器改造为相关性评判器，并根据其与人类判断的一致性以及系统排名对其进行评估。

RQ2 重排序器通过分数阈值化能否泛化为有效的相关性评判器?

为回答**RQ2**，我们采用与**RQ1** 相同的设置，但改用分数阈值化策略（见第 3.2.2 节）。

研究问题 3 基于重排序器的评审器在评估重排序器时，在多大程度上会表现出偏见?

为回答 **研究问题 3**，每个基于重排序器的评审器都会评估其底层的重排序器以及其他重排序器，从而根据某项信息检索评估指标对所有重排序器进行排名；随后，我们分析每个评审器表现出的偏见（见第 3.2.3节）。

3.2 实验设置

3.2.1 不同规模的重排序器。我们使用 3 个重排序器系列：monoT5 [38],RankLLaMA [27]和 Rank1 [57]。monoT5 和 RankLLaMA 在相关文献中得到了广泛应用 [30, 57, 58], ，而 Rank1 [57] 是近期提出的先进基于推理的重排序器。它们都是逐点重排序器：给定一个查询和一个候选文档，它们会分别为其分配相关性分数，最终排名则通过按这些分数对文档排序得到。它们都使用 MS MARCO V1 [6] 的训练集进行训练。

- **monoT5** [38] 对 T5 [41] 进行微调，使其根据文档是否与查询相关，输出两个特殊标记之一：“true”或“false”。monoT5 对“true”和“false”标记的 logits 应用 softmax，并将分配给“true”标记的概率用作相关性分数。
- **RankLLaMA** [27]使用 LoRA [24] 对 Llama 2 [53]进行微调，直接将序列结束标记的表示投影为相关性分数。与 monoT5 不同，RankLLaMA 生成的相关性分数是无界实数分数，而非概率。
- **Rank1** [57] 微调大语言模型，使其先生成推理轨迹，再按照以下输出格式生成两个特殊标记之一

表 1：TREC 2019–2023 深度学习（TREC-DL）赛道的统计数据 [10–14]。对于 TREC-DL 2022 和 2023，我们遵循 [56] 并使用移除自动传播段落后的相关性判断。

赛道	#运行数	#查询	相关性标签 (0/1/2/3)
TREC-DL 2019	37	43	5,158 / 1,601 / 1,804 / 697
TREC-DL 2020	59	54	7,780 / 1,940 / 1,020 / 646
TREC-DL 2021	63	53	4,338 / 3,063 / 2,341 / 1,086
TREC-DL 2022	60	76	12,892 / 6,192 / 3,053 / 1,385
TREC-DL 2023	35	82	11,618 / 3,774 / 1,942 / 1,544

：“<think> ... </think> true/false”。用于训练的推理轨迹由 DeepSeek 的 R1 在 MS MARCO 上生成。与 monoT5 类似，Rank1 将分配给“true”标记的概率作为相关性分数。

我们还研究了重排序器规模对相关性判断预测的影响。我们纳入了 **monoT5-base**（220M）、**monoT5-large**（770M）和**monoT5-3B**；**RankLLaMA-7B** 和**RankLLa-MA-13B**；以及**Rank1-7B**、**Rank1-14B**和 **Rank1-32B**，涵盖了架构和规模各异的 8 种重排序器。

3.2.2 适配方法。 我们设计了两种策略，将重排序器适配为基于重排序器的相关性评判器：

(1) 直接生成 仅适用于 monoT5 [38]和 Rank1 [57]。我们不使用数值相关性分数，而是直接将模型生成的特殊标记映射为相关性标签：如果 monoT5/Rank1 生成“true”，则将文档标记为“相关”；如果生成“false”，则将其标记为不相关。

(2) 分数阈值化 适用于所有重排序器。我们通过应用预定义的阈值，将它们的连续相关性分数转换为二元标签：如果文档的分数大于或等于该阈值，则预测为相关，否则预测为不相关。

本文专注于生成二元相关性判断。许多广泛使用的信息检索评估指标，例如精确率、召回率、平均倒数排名和平均精度均值，都基于二元相关性判断运行，这意味着该设置涵盖了大多数常见的信息检索评估场景。将这些适配策略扩展为支持分级相关性判断是未来工作中一个有趣的方向，但超出了本文的范围。

3.2.3 偏差分析：基于重排序器的评审器评估重排序器。 鉴于第 3.2.1 节所述的 8 个重排序器，我们通过考察每个基于重排序器的评审器如何评估不同的排序系统来进行偏差分析。对于每个评审器，我们整理了 9 个系统运行结果：一个 BM25 检索器 [46], 以及 BM25 后接 8 个重排序器中的每一个。利用该评审器生成的相关性判断，我们计算每个运行结果的标准信息检索评估指标值，并对这 9 个系统进行排名。随后，我们分析这些排名，以刻画每个基于重排序器的评审器所表现出的评估偏差。技术细节请参见第 3.2.7 节。

3.2.4 数据集。 遵循近期研究 [48, 56]，我们使用 TREC 2019–2023 深度学习（TREC-DL）赛道 [10–14]；统计数据见表 1。TREC-DL2019–2020 和 2021–2023 分别基于 MS MARCO V1 和 V2 段落排序集合，包含 880 万和 1.38 亿个段落。遵循 [7], ，我们将 TREC-DL 2019–2020 视为语料库内评估，因为本文使用的所有重排序器均在 MS MARCO v1 上训练；相比之下，TREC-DL 2021–2023 可被视为训练语料库外评估，因为从 TREC-DL 2021 起使用的是 MS MARCO v2 集合。每个数据集都有四个相关性等级：完全相关（3）、高度相关（2）、相关（1）和不相关（0）。我们遵循官方 TREC-DL 论文 [10–14], ，将相关性等级 ≥ 2 视为正例，对段落相关性判断进行二值化。TREC-DL 2019–2021 中的所有相关性标签均由美国国家标准与技术研究院（NIST）的评估员人工标注，而 TREC-DL 2022 和 2023 还包括自动传播的标签，这些标签被分配给近重复段落。遵循 [56],，我们舍弃这些传播标签，仅保留人工标注。

3.2.5 评测指标。 我们从两个角度进行评测：评判一致性和系统排名。对于一致性，遵循先前的工作 [19, 31, 48], ，我们计算人类标注的相关性判断（相关性判断）与基于重排序器的评审器所预测的相关性判断之间的 Cohen κ。对于系统排名，遵循 [55, 56], ，我们计算肯德尔 τ 相关性：该相关性衡量基于 TREC 评估员的判断得出的信息检索评估指标与基于各个重排序器的评审器得出的指标所产生的系统排序之间的相关程度；对于 TREC-DL 数据集，系统排名指该数据集所有已提交运行结果的排序。对于信息检索评估指标，我们遵循先前的研究，使用 MAP@100 [16, 19], 和 MRR@10 [16]。[1]

3.2.6 参考基于大语言模型的评估器。 为了给基于重排序器的相关性评判器提供比较基准，我们使用 UMBRELA [56], 这一最先进的基于大语言模型的相关性评判器作为参考。需要注意的是，本文的目标并不是开发一个能够超越 UMBRELA 的最先进评估器。给定一个查询和一个文档，UM-BRELA 采用零样本描述性、叙事性和方面（DNA）提示方案 [52] 来提示 GPT-4o。

3.2.7 实现细节。对于重排序器检查点，我们使用公开可用的 Hugging Face 检查点，包括 monoT5-base（220M）、monoT5-large（770M）、monoT5-3B（3B）、RankLLa MA-7B、RankLLaMA-13B、Rank1-7B、Rank1-14B 和 Rank1-32B。[2] 对于直接生成（RQ1），我们按照原始实现中的默认解码设置，记录 monoT5 和 Rank1 生成的特殊标记。对于分数阈值化（RQ2），针对每个基于重排序器的评判器，我们选择在一个 TREC-DL 数据集上表现最佳的阈值，并将其应用于另一个数据集，具体如下：TREC-DL 19→20、TREC-DL 20→19、TREC-DL 21→22、TREC-DL 22→21 和 TREC-DL 22→23。对于偏差分析（RQ3），我们使用默认参数 k1=0.9、b=0.4 的 Pyserini BM25 [46] 。[3] BM25 检索排名前 1,000 的文档，然后由各个重排序器对其重新排序。对于 UMBRELA [56], ，我们使用原作者生成的分级相关性判断，并将其二值化；这些判断由作者应请求慷慨提供。所有实验均在配备 RTX 6000 Ada（48GB）GPU 的设备上进行。

[1] 我们还报告了 nDCG@10，结论相似；由于篇幅限制，结果请见 https://github.com/ChuanMeng/reranker-as-judge。[2] 所有检查点均可在 Hugging Face 上公开获取。Castorini（https://huggingface.co/castorini）：monot5-base-msmarco、monot5-large-msmarco、monot5-3b-msmarco、rankllama-v1-7b-lora-passage、rankllama-v1-13b-lora-passage；JHU-CLSP（https://huggingface.co/jhu-clsp）：rank1-7b、rank1-14b、rank1-32b。[3] https://github.com/castorini/pyserini

表2：TREC 评审员与通过直接生成进行适配的各个基于重排序器的相关性评判器之间的相关性判断一致性（Cohen's κ）。此处列出 UMBRELA（最先进的相关性评判器）的结果以供参考。每列中的最佳值以粗体显示，次佳值带下划线。

方法	TREC-DL 19	TREC-DL 20	TREC-DL 21	TREC-DL 22	TREC-DL 23
UMBRELA	**0.499**	**0.450**	**0.492**	0.421	**0.418**
monoT5 基础版	0.389	0.322	0.343	0.239	0.269
monoT5 大型版	0.425	0.323	0.340	0.246	0.291
monoT5 3B	0.386	0.385	0.402	0.279	0.335
Rank1-7B	0.429	**0.450**	0.476	**0.425**	0.382
Rank1-14B	0.459	0.433	0.469	0.421	0.394
Rank1-32B	0.452	0.421	0.455	0.420	0.380

表3：TREC 评审员的相关性判断所产生的系统排序与各个通过直接生成进行适配的基于重排序器的相关性评判器所预测的系统排序之间的肯德尔相关系数。结果展示了基于 MAP@100 和 MRR@10 的系统排序。为供参考，还列出了最先进的相关性评判器 UMBRELA 的结果。每列中的最佳值以粗体显示，次佳值以下划线标出。

方法	TREC-DL 19		TREC-DL 20		TREC-DL 21		TREC-DL 22		TREC-DL 23	
	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10
UMBRELA	**0.946**	0.832	**0.930**	0.837	**0.921**	**0.868**	0.905	**0.774**	0.849	**0.844**
monoT5 基础版	0.883	0.839	0.799	0.807	0.858	0.731	0.810	0.606	0.792	0.479
monoT5 大型版	0.910	**0.841**	0.825	0.702	0.863	0.786	0.825	0.625	0.825	0.741
monoT5 3B	0.895	0.832	0.868	0.692	0.906	0.828	0.879	0.696	0.829	0.669
Rank1-7B	0.901	0.835	0.883	0.821	0.892	0.839	0.922	0.741	0.889	0.817
Rank1-14B	0.892	0.763	0.887	**0.847**	0.890	0.829	**0.937**	0.747	**0.909**	0.797
Rank1-32B	0.916	0.808	0.889	0.834	0.886	0.836	0.927	0.759	0.899	0.834

4 结果与讨论

4.1 通过直接生成让重排序器充当评判器

为回答**RQ1**，我们评估通过直接生成改造的、基于重排序器的相关性评判器，这些评判器基于 monoT5 [38]和 Rank1 [57]；此外，我们还使用 UMBRELA [56], 这一先进的基于大语言模型的相关性评判器作为参照。请注意，直接生成不适用于 RankLLaMA [27], , 因为它会生成连续相关性分数，而不会生成特殊标记（例如"true"或"false"）。结果见表2 （以 Cohen's κ衡量的相关性判断一致性）和表3 （以 MAP@100 和 MRR@10 衡量的系统排名）。我们有三个主要观察结果。

首先，基于 monoT5 和 Rank1 的评判器与 TREC 评审员的一致性（Cohen's κ）不低于 UMBRELA 的情况约占 40%；此外，其系统排名相关性不低于 UMBRELA 的情况约占 50%。例如，如表 2所示，Rank1-7B 在 TREC-DL 20 上取得了与 UMBRELA 相同的 κ 值（0.450），在 TREC-DL 22 上则取得了更高的值（0.425）。同样，如表 3所示，monoT5-large 在 TREC-DL 19 上的 MRR@10 肯德尔相关系数最高（0.841）；而 Rank1-14B 在 TREC-DL 20 上的 MRR@10（0.847）以及 TREC-DL 22 上的 MAP@100（0.937）相关系数最高；在 TREC-DL 23 上，Rank1-14B 的 MAP@100 相关系数再次最高（0.909）。

其次，基于 Rank1 的评判器通常优于基于 monoT5 的评判器，并且在不同数据集上表现出更强的稳健性。具体而言，monoT5 在 TREC-DL 22 和 23 上的性能明显下降，而 Rank1 则相对稳定。例如，monoT5 在 TREC-DL 22 上的大多数 Cohen's κ 值都低于 0.3，而在 TREC-DL 23 上也如此；而 Rank1-7B 在 TREC-DL 22 上的一致性最高（0.425），Rank1-14B 在 TREC-DL 23 上的一致性排名第二（0.394）。请注意，所有重排序器都基于 MS MARCO v1 段落排序语料库 [6] 进行训练。TREC-DL 19–20 基于 MS MARCO v1，而 TREC-DL 21–23 采用 MS MARCO v2 段落排序语料库。这些观察结果表明，基于推理的 Rank1 在面对语料库变化时表现出更强的稳健性。

第三，扩大重排序器的规模会产生喜忧参半的结果：通常能提升系统排序性能，但并不总能提升相关性判断一致性。例如，在大多数情况下，Rank1-14B 和 Rank1-32B 的肯德尔 τ 相关性高于 Rank1-7B；然而，从 14B 扩大到 32B 并未带来明显优势。相比之下，在 60% 的情况下，Rank1-7B 的相关性判断一致性高于 Rank1-14B 和 Rank1-32B。

4.2 通过分数阈值化将重排序器用作评判器

为回答 **RQ2**，我们评估了通过分数阈值化改造的基于重排序器的相关性评判器，这些评判器改编自 monoT5 [38], RankLLaMA [27], 和 Rank1 [57]；我们还将 UMBRELA [56], 这一最先进的基于大语言模型的相关性评判器作为参照。阈值选择的详细信息见第 3.2.7 节。我们在表4 中报告结果（以 Cohen's κ衡量的相关性判断一致性），并在表 5 中报告结果（以 MAP@100 和 MRR@10 衡量的系统排名）。

与 **RQ1** 的结果类似，基于重排序器的评审器在相关性判断一致性和系统排名相关性两方面，约 40% 的情况下表现优于 UMBRELA。基于 Rank1 的评判器依然比基于 monoT5 的评判器表现更强、更稳健。扩展

表4：TREC 评审员与通过分数阈值化调整的各个基于重排序器的相关性评判器之间的相关性判断一致性（Cohen's κ）。此处列出 UMBRELA（最先进的相关性评判器）的结果以供参考。每列中的最佳值以粗体标出，次佳值加下划线。

方法	TREC-DL 19	TREC-DL 20	TREC-DL 21	TREC-DL 22	TREC-DL 23
UMBRELA	**0.499**	0.450	**0.492**	0.421	**0.418**
monoT5 基础版	0.350	0.307	0.341	0.240	0.269
monoT5 大型版	0.398	0.311	0.373	0.256	0.295
monoT5 3B	0.391	0.364	0.273	0.273	0.336
RankLLaMA-7B	0.400	0.371	0.387	0.232	0.311
RankLLaMA-13B	0.407	0.340	0.360	0.260	0.314
Rank1-7B	0.443	**0.455**	0.482	**0.430**	0.386
Rank1-14B	0.463	0.438	0.472	0.425	0.406
Rank1-32B	0.454	0.417	0.461	0.426	0.392

表 5：TREC 评审员的相关性判断所产生的系统排序与各个经分数阈值化调整的基于重排序器的相关性评判器所预测的系统排序之间的肯德尔 τ 相关系数。结果展示了基于 MAP@100 和 MRR@10 的系统排序。此处列出最先进的相关性评判器 UMBRELA 的结果以供参考。每列中的最佳值以粗体显示，次佳值带下划线。

方法	TREC-DL 19		TREC-DL 20		TREC-DL 21		TREC-DL 22		TREC-DL 23	
	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10
UMBRELA	**0.946**	0.832	**0.930**	**0.837**	**0.921**	**0.868**	0.905	**0.774**	0.849	**0.844**
monoT5 基础版	0.889	**0.846**	0.795	0.668	0.872	0.784	0.776	0.570	0.808	0.487
monoT5 大型版	0.898	0.749	0.832	0.714	0.870	0.814	0.833	0.668	0.815	0.669
monoT5 3B	0.892	0.818	0.868	0.752	0.902	0.842	0.887	0.727	0.835	0.719
RankLLaMA-7B	0.889	0.806	0.877	0.664	0.860	0.799	0.834	0.660	0.788	0.655
RankLLaMA-13B	0.892	0.781	0.876	0.651	0.856	0.836	0.841	0.768	0.785	0.700
Rank1-7B	0.910	0.835	0.884	0.795	0.881	0.844	0.919	0.725	0.886	0.804
Rank1-14B	0.898	0.763	0.887	0.820	0.889	0.822	0.924	0.746	**0.906**	0.797
Rank1-32B	0.910	0.808	0.890	0.830	0.882	0.835	**0.925**	0.751	0.896	**0.844**

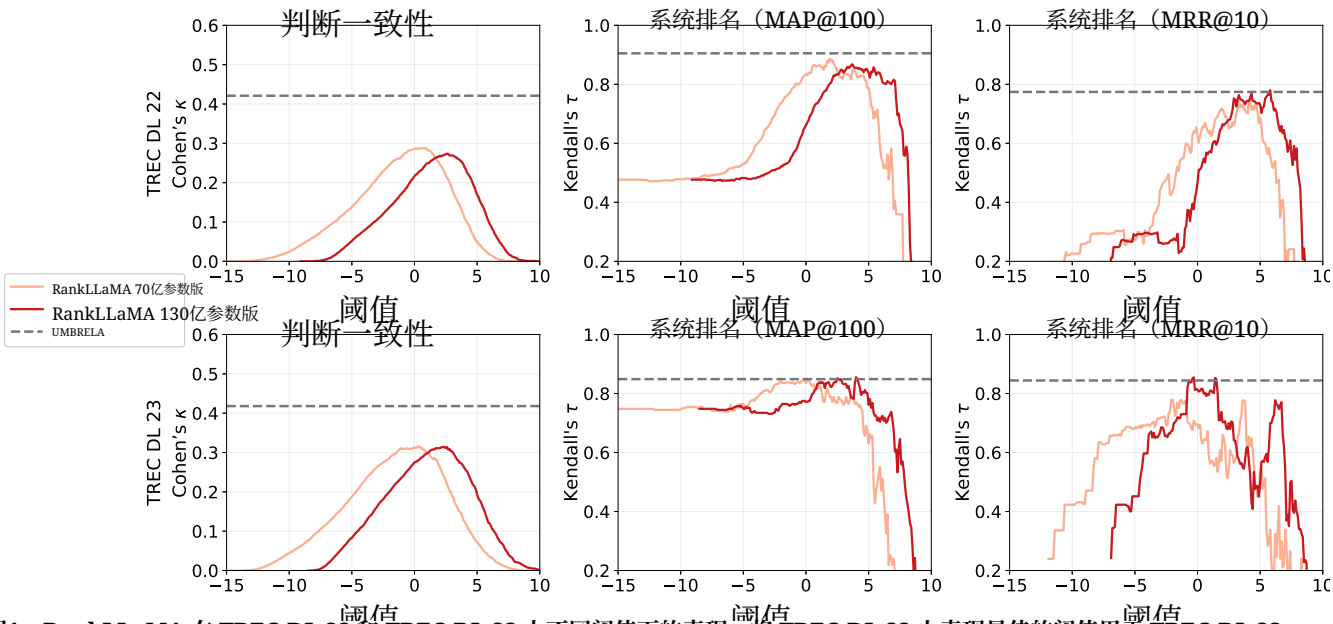

图1：RankLLaMA 在 TREC-DL 22 和 TREC-DL 23 上不同阈值下的表现。将 TREC-DL 22 上表现最佳的阈值用于 TREC-DL 23。

重排序器再次带来喜忧参半的结果，且对系统排序性能的提升往往大于对相关性判断一致性的提升。

与RQ1中的发现不同，在分数阈值化下，将 Rank1 扩展至 32B 比 14B 和 7B 变体带来更多收益。

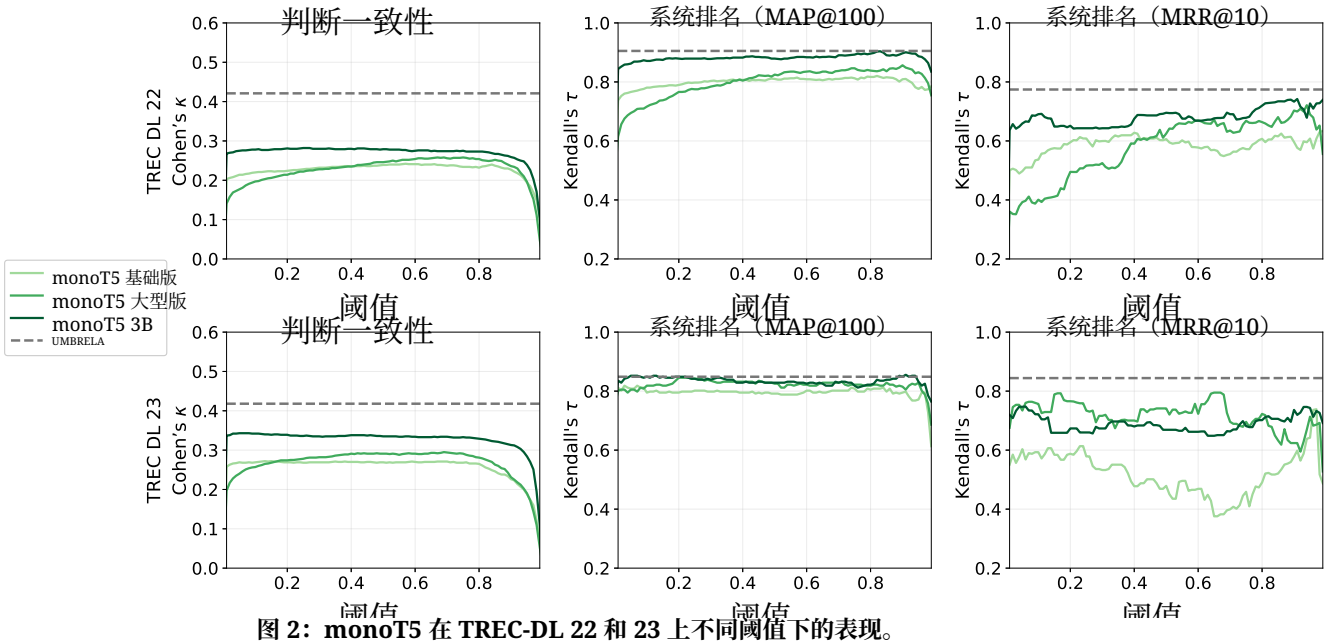

图 2：monoT5 在 TREC-DL 22 和 23 上不同阈值下的表现。

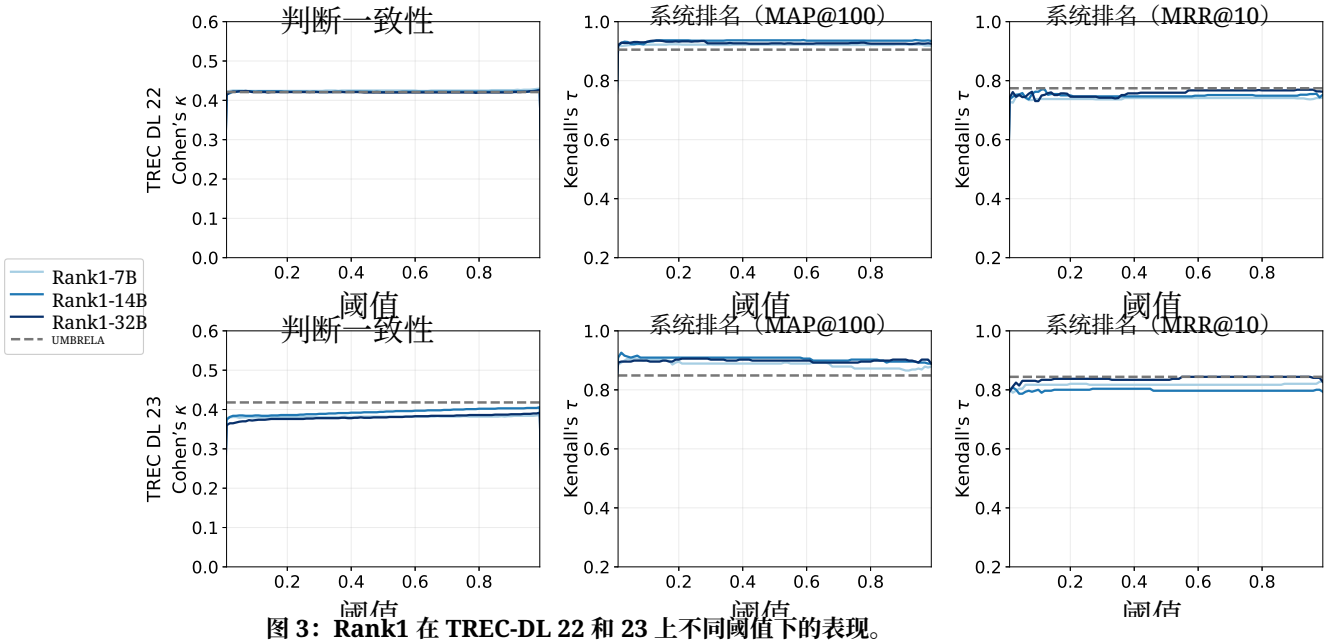

图 3：Rank1 在 TREC-DL 22 和 23 上不同阈值下的表现。

我们还观察到，基于 RankLLaMA 的评判器通常不如基于 Rank1 的评判器，且相比基于 monoT5 的评判器也没有明显优势。为进一步探究这一现象，我们分析了 RankLLaMA、monoT5 和 Rank1 在 TREC-DL 22 和 23 上不同阈值下的表现，如图 1、2 和 3 所示。请注意，我们将 TREC-DL 22 上表现最佳的阈值应用于 TREC-DL 23（详情见第 3.2.7 节）。由此得出两项关键发现。首先，对于系统排序性能而言，RankLLaMA 对阈值选择十分敏感，尤其是在 MRR@10 方面。例如，如图1所示，RankLLaMA-13B 在 TREC-DL 22 上以 MRR@10 衡量时的最优阈值约为 5，但随后发生偏移在 TREC-DL 23 上约为 0。就 MRR@10 而言，monoT5 也表现出类似程度的阈值敏感性。

其次，如图3所示，与 RankLLaMA 和 monoT5 相比，Rank1 对阈值变化表现出显著更强的鲁棒性。我们进一步仔细考察发现：Rank1 预测的相关性分数（范围为 $[0, 1]$）高度集中在两端：分数通常接近 1（例如 0.99）或接近 0。相比之下，monoT5 产生的相关性分数中，落在 $[0, 1]$ 范围中间部分的比例更高。因此，基于 Rank1 的评判器对阈值选择的敏感度显著更低。结果表明，在采用分数阈值化时，高度集中的相关性分数能够在不同数据集之间实现更强的泛化鲁棒性。

表 6：BM25 以及结合 3 个重排序器系列中的 8 个重排序器（monoT5（220M、770M 和 3B）、RankLLaMA（7B 和 13B）以及 Rank1（7B、14B 和 32B））的排序性能。BM25 检索排名前 1,000 的条目，随后由各个重排序器重新排序。结果以 MRR@10 和 MAP@100 表示。我们还报告 Judge@10 [40],，该指标衡量排名前 10 的文档中经人工标注员评判的条目所占比例。

方法	TREC-DL 19			TREC-DL 20			TREC-DL 21		
	MAP@100	平均倒数排名	Judge@10	MAP@100	平均倒数排名	Judge@10	MAP@100	平均倒数排名	Judge@10
BM25	0.248	0.702	1.000	0.269	0.653	0.994	0.136	0.498	1.000
BM25+monoT5 基础版	0.430	0.871	0.949	0.478	0.880	0.987	0.309	0.824	0.955
BM25+monoT5 大型版	0.439	0.873	0.949	0.479	0.831	0.976	0.313	0.833	0.955
BM25+monoT5 3B	0.437	0.867	0.965	0.522	0.886	0.989	0.306	0.793	0.957
BM25+RankLLaMA-7B	0.457	0.841	0.958	0.521	0.906	0.985	0.339	0.825	0.972
BM25+RankLLaMA-13B	0.451	0.895	0.977	0.512	0.861	0.978	0.342	0.851	0.972
BM25+Rank1-7B	0.402	0.810	0.847	0.450	0.790	0.824	0.272	0.778	0.757
BM25+Rank1-14B	0.399	0.754	0.809	0.453	0.736	0.856	0.263	0.735	0.713
BM25+Rank1-32B	0.416	0.828	0.830	0.467	0.774	0.880	0.289	0.750	0.777

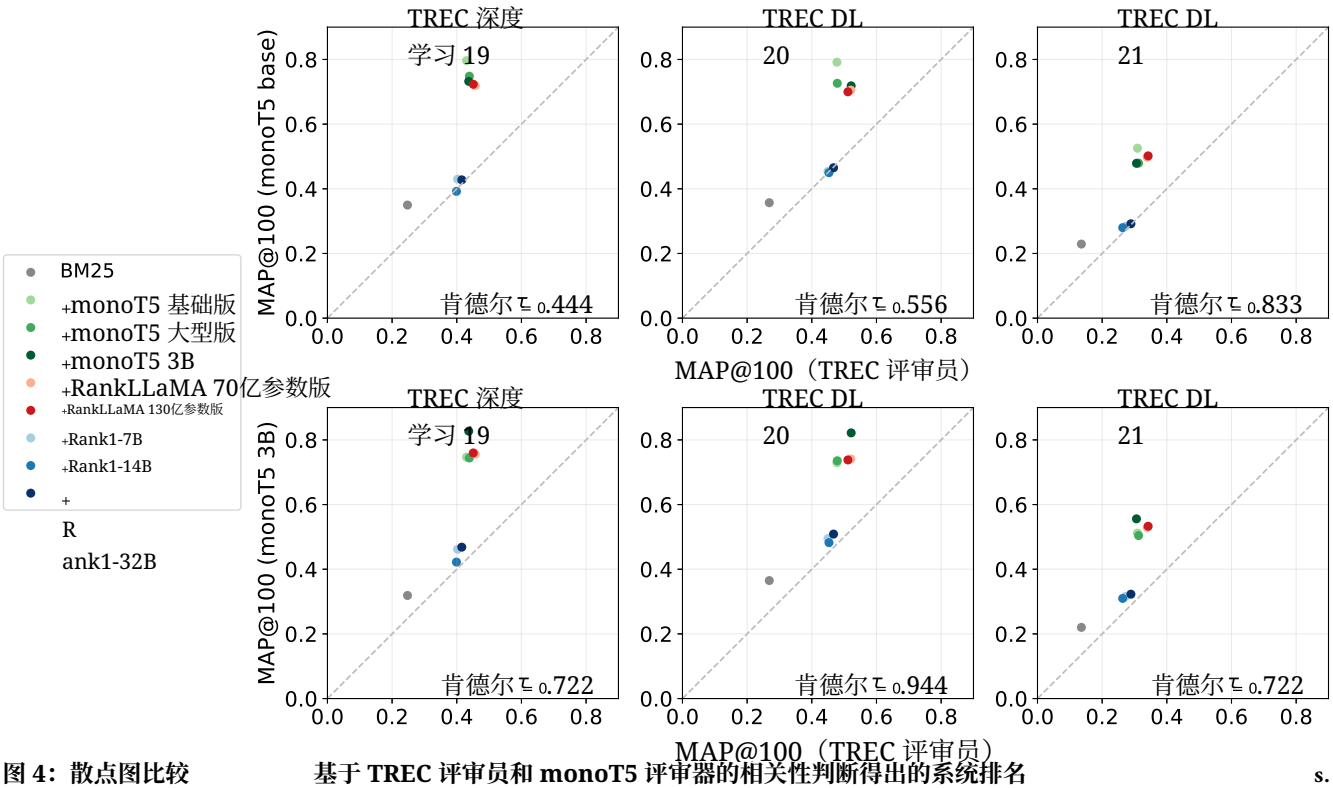

图 4：散点图比较　　基于 TREC 评审员和 monoT5 评审器的相关性判断得出的系统排名　　s.

4.3 基于重排序器的评审器的偏差

为回答**研究问题 3**，我们为每个 TREC-DL 数据集整理了 9 个运行文件：BM25，以及由 BM25 生成的结果，还有 8 个重排序器的结果：monoT5（220M、770M、3B）、RankLLaMA（7B、13B）和 Rank1（7B、14B、32B）。详情请参阅第 3.2.3 节和3.2.7 节。每个基于重排序器的评审器均用于评估所有系统并对其排名，包括重排序后的运行结果和 BM25 运行结果，评估指标为 MAP@100。

作为合理性检查，我们首先在表6中报告使用 TREC 评审员标注的相关性判断，在 TREC-DL 19、20 和 21 上评估的 9 个运行结果的排序性能。我们观察到，尽管 Rank1 在**RQ1/RQ2**中被证明是有效的评审器，但在 TREC 评审员标注的相关性判断 qrels 下，其重排序质量有限。例如，BM25 402, 0 1 7 100 0 450, 0 272 19, 20, 21, –Rank - B 在 TREC-DL 和相应数据集上取得的平均精度均值@分数分别为 . . 和 。其 Judge@10 值（0.847、0.824、0.757）明显低于相同数据集上 BM25– RankLLaMA-7B 的值（0.958、0.985、0.972）。这可能是因为 Rank1 返回的许多相关文档未在 TREC 相关性判断中进行评判，导致性能被低估。我们的发现与 [57] 一致。

随后，我们使用散点图将基于 TREC 评审员相关性判断得出的系统排名，与基于重排序器的评审器预测的系统排名进行比较，以此对结果进行可视化。由于篇幅所限，我们展示 TREC-DL 19、20 和 21 上的结果，其中评审器分别基于 monoT5-base/3B、RankLLaMA-7B/13B 以及 Rank1-7B/32B。

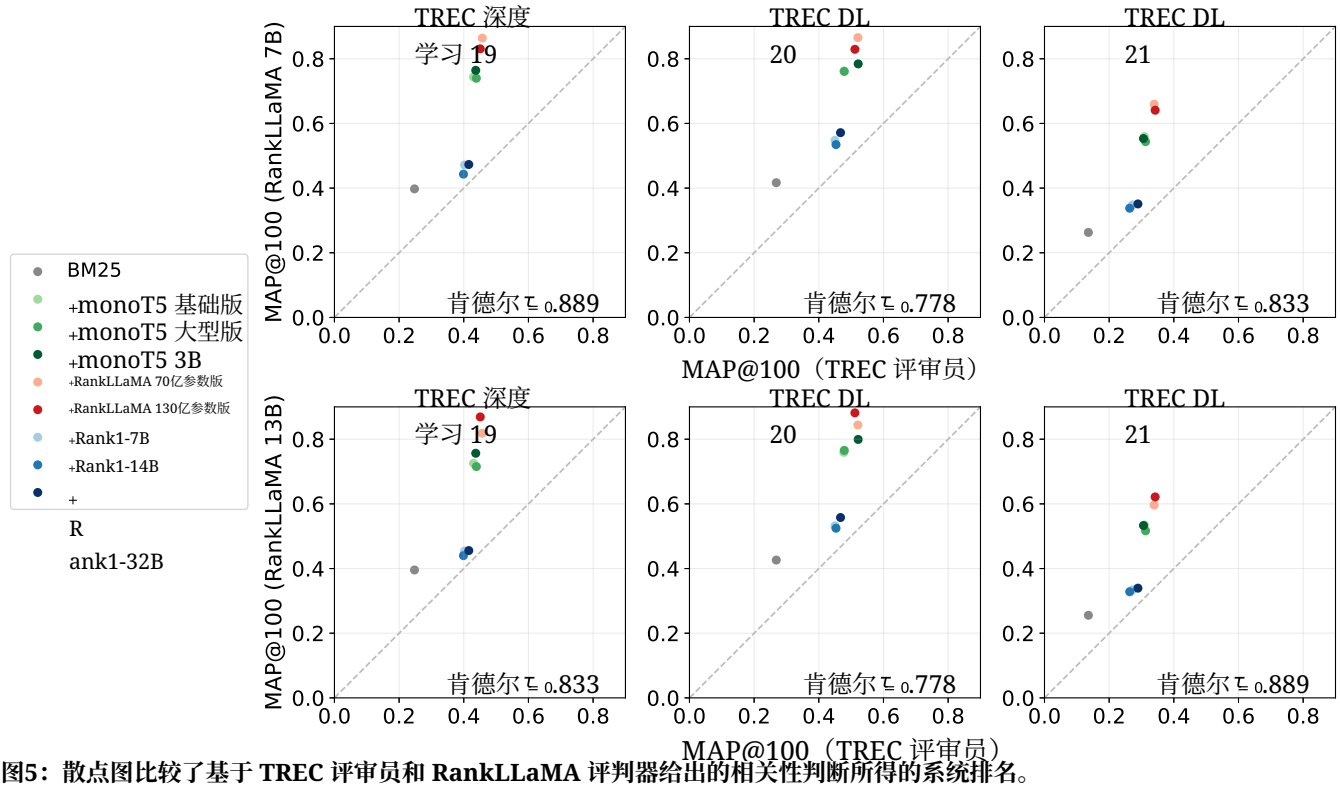

图5：散点图比较了基于 TREC 评审员和 RankLLaMA 评判器给出的相关性判断所得的系统排名。

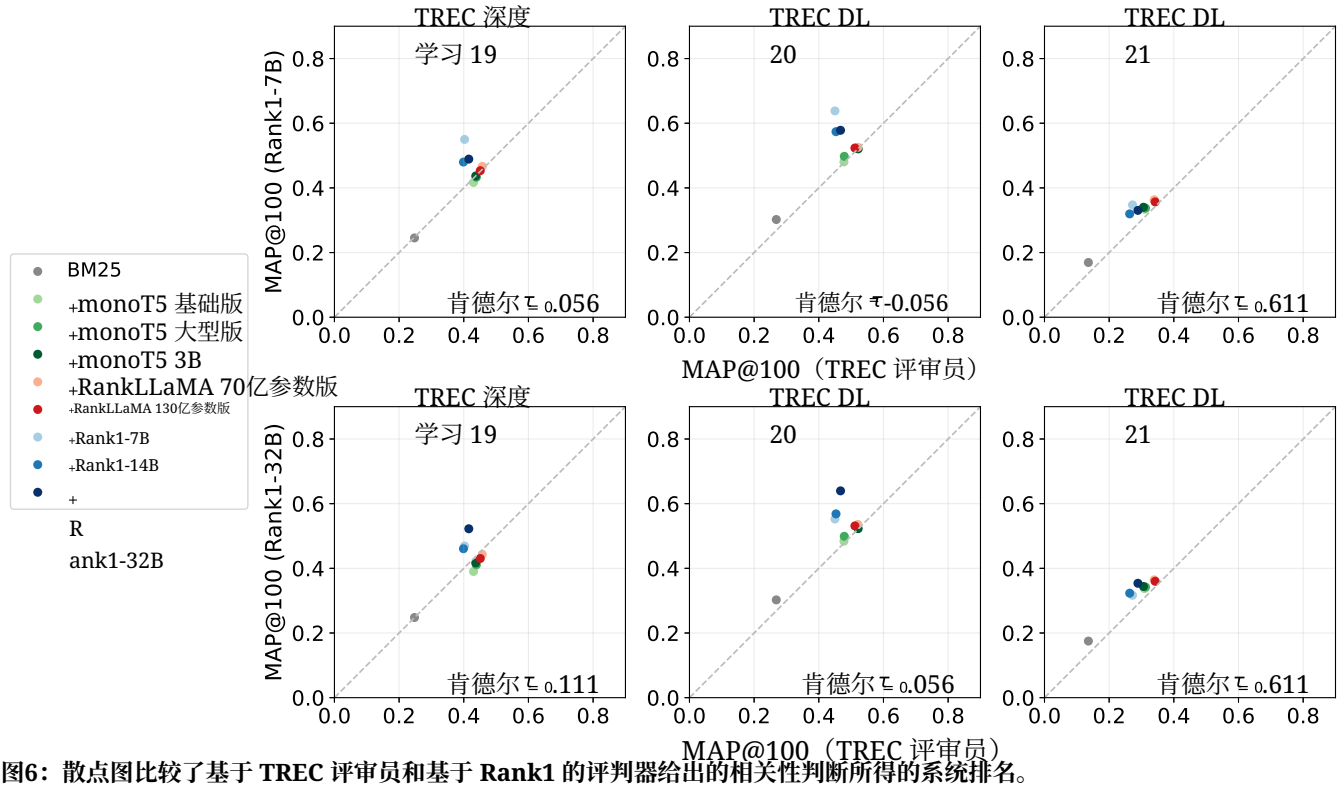

图6：散点图比较了基于 TREC 评审员和基于 Rank1 的评判器给出的相关性判断所得的系统排名。

在图4、5和6中。请注意，monoT5 和 Rank1 使用直接生成（RQ1），而 RankLLaMA 使用阈值

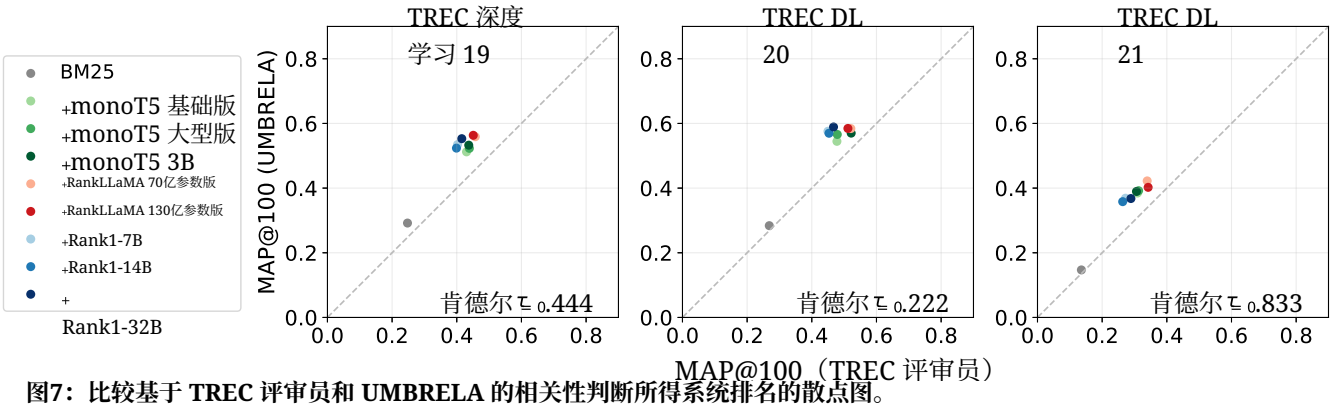

图7：比较基于 TREC 评审员和 UMBRELA 的相关性判断所得系统排名的散点图。

见 RQ2。图7中还展示了 UMBRELA 作为参考评判者。我们有三个主要观察结果。

首先，所有基于重排序器的评审器都表现出明显的自我偏好：每个评审器倾向于给自己的重排序器以及同一家族中的其他重排序器打出最高分。例如，在图 5中，RankLLaMA-7B 评审器将 RankLLaMA-7B 排在第一位，将 RankLLaMA-13B 排在第二位；而 RankLLaMA-13B 评审器则呈现相反的模式。

其次，monoT5 和 RankLLaMA 评判器都倾向于将彼此评为仅次于自身家族的第二佳家族。它们始终给 Rank1 的评分低得多，有时甚至认为其重排序表现与单独使用 BM25 相近。这表明存在跨家族偏差：非推理型重排序器更偏向其他非推理型家族，而非推理型 Rank1 系列。

第三，UMBRELA 和 Rank1 能准确预测 BM25 的表现（BM25 的灰色点靠近对角线），而基于 monoT5 和 RankLLaMA 的评判器则倾向于高估其表现。

5 相关工作

近期关于使用大语言模型作为相关性评判器的研究，大致可以分为三个方向。第一，大量研究聚焦于预测相关性判断的方法，包括提示设计 [15, 19, 20, 52], 在人工标注数据上微调大语言模型 [2, 23, 31], 上下文示例选择 [29], 流水线设计 [48], 文档摘要 [37], 以及集成多个基于大语言模型的评估器 [44]。第二，许多研究分析基于大语言模型的评估器的行为，例如考察认知偏差 [8],对顶尖系统之间排名差异的保留程度 [39],以及对输入变化的敏感性 [3, 5]。第三，一些研究探索实际应用，包括补全缺失的相关性判断 [1, 2, 23, 54],构建完全由合成数据生成的测试集合 [42, 43], 以及在预算有限的情况下，结合人工标注和大语言模型生成的标注来构建测试集合 [51]。

最密切相关的工作包括 [9, 49], [22], 和 [28]。[9, 49] 讨论了排名与相关性判断之间的概念联系；相比之下，我们通过复现重排序器作为相关性评判器，开展了系统的实证研究。 [22] 分析了评估偏差，并表明基于大语言模型的评判器（按照 UMBRELA [56]构建）偏向基于大语言模型的搜索系统；相比之下，我们的工作聚焦于将重排序器用作相关性评判器。 [28] 在单次提示设置中使用成对重排序器，并为每个查询提供一个已知的相关文档；相比之下，我们研究的是在更现实的设置中、不依赖先验相关性信息的逐点重排序器。

6 结论与未来工作

在本研究中，我们复现了来自三个家族、规模各异的 8 个重排序器，并将其用作相关性评判器。为此，我们设计了两种适配策略——直接生成和分数阈值化，使重排序器能够充当相关性评判器。此外，我们还分析了基于重排序器的评审器所表现出的评估偏差。总体而言，在 TREC-DL 2019–2023 上进行的实验表明，大多数关于重排序的发现也适用于“重排序器作为相关性评判器”的设置。具体来说，我们观察到：(i) 在这两种策略下，基于重排序器的评审器均表现出有竞争力的性能，在约 40%–50% 的情况下优于 UMBRELA 这一最先进的基于大语言模型的评估器。在基于重排序器的评审器中，基于推理的 Rank1 总体表现最佳，并且在不同数据集上的稳健性更强。扩大重排序器的模型规模通常有益，但收益并不总是稳定。(ii) 对相关性分数高度集中在两端附近的重排序器而言，分数阈值化尤其有效；Rank1 预测的相关性分数往往非常接近 1 或接近 0，因此与其他重排序器相比，它对阈值选择不那么敏感。(iii) 基于重排序器的评审器表现出明显的偏差：偏向自身及同一家族的重排序器，存在跨家族偏差，并且往往会高估 BM25 的检索质量。

我们的研究进一步表明，相关性判断并非一项本质上全新的任务，而是相关性预测的一种特定情形，这与先前的论述一致 [9, 49]。这一联系意味着，我们的领域可以基于成熟的相关性预测方法（例如重排序器）来构建相关性评判器，而非从头设计新的相关性评判器。通过重用现有模型和数据，我们可以减少重复劳动，避免现有资源未得到充分利用 [25]。

未来的工作包括：（i）缓解基于重排序器的评审器中的偏差；受近期研究启发， [23], 通过添加适配器参数，将重排序器适配为相关性评判器，似乎是一个很有前景的方向。（ii）虽然许多广泛使用的信息检索指标（例如精确率、召回率、平均倒数排名和平均精度均值）采用二元判定，但提出支持分级标签的适配策略是顺理成章的下一步。（iii）将重排序器适配到逐点评分之外的形式（例如列表式 [26, 45, 58]），并使用其他包含高推理强度查询（例如 BRIGHT [50]）或对话式查询 [32, 34–36]的数据集，也将很有价值。（iv）Rank1 的表现可能被低估（见表6），这凸显了我们的领域有必要重新审查新排序器返回的未判定文档。

致谢

本研究获得了一项图灵人工智能加速奖学金的支持，该奖学金由工程与自然科学研究委员会（EPSRC）资助，资助编号为 EP/V025708/1。本研究还获得了荷兰研究委员会（NWO）的支持，项目编号为 024.004.022、NWA.1389.20.183 和 KICH3.LTP.20.006；此外还获得了欧盟的支持，资助协议编号为 101070212 （ FINDHR ）和 101201510 （ UNITE ）。

所有内容均代表作者的观点，不一定为其各自的雇主和/或赞助方所认同或认可。

参考文献

[1] Zahra Abbasiantaeb、Chuan Meng、Leif Azzopardi 和 Mohammad Aliannejadi。2024年。我们能否利用大语言模型填补相关性判断空缺？ *arXiv* 预印本 *arXiv:2405.05600* (2024)。[2] Zahra Abbasiantaeb、Chuan Meng、Leif Azzopardi 和 Mohammad Aliannejadi。2025年。提升对话式搜索测试集合的可复用性。载于*ECIR*。196–213。[3] Marwah Alaofi、Paul Thomas、Falk Scholer 和 Mark Sanderson。2024年。大语言模型可能被诱导将某文档标注为相关：附近最好的咖啡馆；这篇论文完全相关。载于*SIGIR-AP*。32–41。[4] Negar Arabzadeh 和 Charles LA Clarke。2025年。大语言模型相关性判断方法的基准测试。载于*SIGIR*。3194–3204。[5] Negar Arabzadeh 和 Charles LA Clarke。2025年。对大语言模型相关性判断中提示敏感性的人工智能比较分析。载于*SIGIR*。2784–2788。[6] Payal Bajaj、Daniel Campos、Nick Craswell、Li Deng、Jianfeng Gao、Xiaodong Liu、Rangan Majumder、Andrew McNamara、Bhaskar Mitra、Tri Nguyen、Mir Rosenberg、Xia Song、Alina Stoica、Saurabh Tiwary 和 Tong Wang。2018年。MS MARCO：人工生成的机器阅读理解数据集。*arXiv* 预印本 *arXiv:1611.09268* (2018)。[7] Leonid Boytsov、David Akinpelu、Nipun Katyal、Tianyi Lin、Fangwei Gao、Yutian Zhao、Jeffrey Huang 和 Eric Nyberg。2025年。长文档排序中的位置偏差：影响、评估与缓解。*arXiv* 预印本 *arXiv:2207.01262* (2025)。[8] Nuo Chen、Jiqun Liu、Xiaoyu Dong、Qijiong Liu、Tetsuya Sakai 和 Xiao-Ming Wu。2024年。人工智能也会产生认知偏差：大语言模型批量相关性评估中的阈值启动探索性研究。载于*SIGIR-AP*。54–63。[9] Charles LA Clarke 和 Laura Dietz。2024年。基于大语言模型的相关性评估仍无法取代人工相关性评估。*arXiv* 预印本 *arXiv:2412.17156* (2024)。[10] Nick Craswell、Bhaskar Mitra、Emine Yilmaz 和 Daniel Campos。2020年。TREC 2020 深度学习赛道概述。载于*TREC 2020*。[11] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Campos 和 Ellen M. Voorhees。2019年。TREC 2019 深度学习赛道概述。载于*REC 2019*。[12] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Fernando Campos 和 Jimmy Lin。2021年。TREC 2021 深度学习赛道概述。载于*TREC 2021*。[13] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Fernando Campos、Jimmy Lin、Ellen M. Voorhees 和 Ian Soboroff。2022年。TREC 2022 深度学习赛道概述。载于*TREC 2022*。[14] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Hossein A. Rahmani、Daniel Campos、Jimmy Lin、Ellen M. Voorhees 和 Ian Soboroff。2023年。TREC 2023 深度学习赛道概述。载于*TREC 2023*。[15] Mouly Dewan、Jiqun Liu 和 Chirag Shah。2025年。面向信息检索评估的、由大语言模型驱动的有用性标注。载于第*48*届 *ACM* 信息检索研究与开发国际 *SIGIR* 会议论文集。3055–3059。[16] Mouly Dewan、Jiqun Liu 和 Chirag Shah。2025年。TRUE：用于信息检索中大语言模型驱动相关性判断的可复现框架。*arXiv* 预印本 *arXiv:2509.25602* (2025)。[17] Laura Dietz、Oleg Zendel、Peter Bailey、Charles Clarke、Ellese Cotterill、Jeff Dalton、Faegheh Hasibi、Mark Sanderson 和 Nick Craswell。2025年。大语言模型评测的常见套路：关于大语言模型评测有效性的观点。*arXiv* 预印本 *arXiv:2504.19076* (2025)。[18] Laura Dietz、Oleg Zendel、Peter Bailey、Charles LA Clarke、Ellese Cotterill、Jeff Dalton、Faegheh Hasibi、Mark Sanderson 和 Nick Craswell。2025年。使用大语言模型评审器的原则与指南。载于*ICTIR*。218–229。[19] Guglielmo Faggioli、Laura Dietz、Charles LA Clarke、Gianluca Demartini、Matthias Hagen、Claudia Hauff、Noriko Kando、Evangelos Kanoulas、Martin Potthast、Benno Stein 等人。2023年。关于大语言模型用于相关性判断的观点。载于*ICTIR*。39–50。[20] Naghmeh Farzi 和 Laura Dietz。2025年。基于标准的大语言模型相关性判断。载于*ICTIR*。254–263。[21] Naghmeh Farzi 和 Laura Dietz。2025年。UMBRELA 能否适用于其他大语言模型？收录于 SIGIR。3214–3222。[22] Maik Fröbe、Andrew Parry、Ferdinand Schlatt、Sean MacAvaney、Benno Stein、Martin Potthast 和 Matthias Hagen。2025年。大语言模型相关性评估者彼此之间的共识高于他们与人工评估者之间的共识。收录于 SIGIR。2858–2863。[23] Lukas Gienapp、Martin Potthast、Harrisen Scells 和 Eugene Yang。2025年。特定主题的分类器比提示式大语言模型更适合作为相关性评判器。arXiv 预印本 arXiv:2510.04633（2025年）。[24] Edward J Hu、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。2021年。LoRA：大语言模型的低秩适配。收录于 ICLR。[25] Tianji Jiang、Wenqi Li 和 Jiqun Liu。2025年。交互式信息检索中的数据复用图景：动机、来源与可复用性评估。《美国信息科学与技术学会会刊》76卷第9期，2025年，1258–1276页。[26] Wenhan Liu、Xinyu Ma、Weiwei Sun、Yutao Zhu、Yuchen Li、Dawei Yin 和 Zhicheng Dou。2025年。ReasonRank：以强大的推理能力提升段落排序。arXiv 预印本 arXiv:2508.07050（2025年）。[27] Xueguang Ma、Liang Wang、Nan Yang、Furu Wei 和 Jimmy Lin。2024年。微调 LLaMA 以实现多阶段文本检索。收录于 SIGIR。2421–2425。[28] Sean MacAvaney 和 Luca Soldaini。2023年。用于自动相关性估计的单次标注。收录于 SIGIR。2230–2235。[29] Jack McKechnie、Graham McDonald 和 Craig Macdonald。2025年。为大语言模型生成的相关性评估选择上下文示例。收录于 ECIR。293–309。[30] Chuan Meng、Negar Arabzadeh、Arian Askari、Mohammad Aliannejadi 和 Maarten de Rijke。2024年。面向基于大语言模型的重排序的排序列表截断。收录于 SIGIR。141–151。[31] Chuan Meng、Negar Arabzadeh、Arian Askari、Mohammad Aliannejadi 和 Maarten de Rijke。2025年。使用大语言模型生成的相关性判断进行查询性能预测。TOIS（2025年）。[32] Chuan Meng、Francesco Tonolini、Fengran Mo、Nikolaos Aletras、Emine Yilmaz 和 Gabriella Kazai。2025年。弥合差距：从对话中的即席搜索到主动搜索。收录于 SIGIR。64–74。[33] Si mone Merlo、Stefano Marchesin、Guglielmo Faggioli 和 Nicola Ferro。2025年。评估大语言模型生成的相关性判断的经济高效框架。收录于 CIKM。2115–2126。[34] Fengran Mo、Yifan Gao、Chuan Meng、Xin Liu、Zhuofeng Wu、Kelong Mao、Zhengyang Wang、Pei Chen、Zheng Li、Xian Li 等。2025年。UniConv：统一对话场景下大语言模型的检索与响应生成。收录于 ACL。6936–6949。[35] Fengran Mo 、 Chuan Meng 、 Mohammad Aliannejadi 和 Jian-Yun Nie。2025年。对话式搜索：从基础到大语言模型时代的前沿。收录于 SIGIR。4094–4097。[36] Fengran Mo、Jinghan Zhang、Yuchen Hui、Jia Ao Sun、Zhichao Xu、Zhan Su 和 Jian-Yun Nie。2025年。Conv Mix：面向对话式稠密检索的混合标准数据增强框架。arXiv 预印本 arXiv:2508.04001（2025年）。[37] Samaneh Mohtadi、Kevin Roitero、Stefano Mizzaro 和 Gianluca Demartini。2025年。文档摘要对基于大语言模型的相关性判断的影响。arXiv 预印本 arXiv:2512.05334（2025年）。[38] Rodrigo Nogueira、Zhiying Jiang、Ronak Pradeep 和 Jimmy Lin。2020年。使用预训练序列到序列模型进行文档排序。收录于 EMNLP。708–718。[39] David Otero、Javier Parapar 和 Álvaro Barreiro。2025年。使用大语言模型进行自动相关性评估以实现公平可靠的检索评估：局限性。收录于 SIGIR。2545–2549。[40] Ronak Pradeep、Sahel Sharifymoghaddam 和 Jimmy Lin。2023年。RankZephyr：高效、稳健的零样本列表式重排序轻而易举！arXiv 预印本 arXiv:2312.02724（2023年）。[41] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2020年。探索统一文本到文本转换器的迁移学习极限。《机器学习研究期刊》21卷140期（2020年），1–67页。[42] Hossein A. Rahmani、Nick Craswell、Emine Yilmaz、Bhaskar Mitra 和 Daniel Ca mpos。2024年。用于检索评估的合成测试集。收录于 SIGIR。2647–2651。[43] Hossein A. Rahmani、Xi Wang、Emine Yilmaz、Nick Craswell、Bhaskar Mitra 和 Paul Thomas。2025年。SynDL：用于段落检索的大规模合成测试集。收录于 WWW。781–784。[44] Hossein A. Rahmani、Emine Yilmaz、Nick Craswell 和 Bhaskar Mitra。2025年。JudgeBlender：集成自动相关性判断。收录于 WWW。1268–1272。[45] Mandeep Rathee、Sean MacAvaney 和 Avishek Anand。2025年。使用基于大语言模型的列表式排序器引导检索。收录于 ECIR。230–246。[46] Stephen E. Robertson 、 Steve Walker 、 Susan Jones 、 Micheline M Hancock-Beaulieu、Mike Gatford 等。1995年。TREC-3 中的 Okapi。NIST 特别出版物 109（1995年），109。

[47] Tetsuya Sakai、Khant Myoe Rain、Rikiya Takehi、Sijie Tao 和 Young-In Song。2025年。开源基于大语言模型的相关性评估与高度可靠的人工相关性评估：案例研究。收录于 *CIKM*。5186–5190。[48] Julian A. Schnabel、Johanne R. Trippas、Falk Scholer 和 Danula Hettiachchi。2025年。多阶段大语言模型流水线在相关性评估中可胜过 GPT-4o。收录于 *WWW*。1288–1292。[49] Ian Soboroff。2025年。不要使用大语言模型进行相关性判断。*Information Retrieval Research* 1卷1期（2025年），29–46页。[50] Hongjin Su、Howard Yen、Mengzhou Xia、Weijia Shi、Niklas Muennighoff、Han-yu Wang、Liu Haisu、Quan Shi、Zachary S. Siegel、Michael Tang 等。2025年。BRIGHT：用于推理密集型检索的真实且具有挑战性的基准测试。收录于 *ICLR*。[51] Rikiya Takehi、Ellen M Voorhees、Tetsuya Sakai 和 Ian Soboroff。2025年。大语言模型辅助的相关性评估：何时应寻求大语言模型的帮助？收录于 *SIGIR*。95–105。[52] Paul Thomas、Seth Spielman、Nick Craswell 和 Bhaskar Mitra。2024年。大语言模型能够准确预测搜索者偏好。收录于 *SIGIR*。1930–1940。

[53] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yas- mine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhos- ale 等。2023年。Llama 2：开放式基础模型与微调聊天模型。*arXiv* 预印本 *arXiv:2307.09288*（2023年）。[54] Shivani Upadhyay、Ehsan Kamalloo 和 Jimmy Lin。2024年。大语言模型可以弥补评测中的相关性判断缺失。*arXiv* 预印本 *arXiv:2405.04727*（2024年）。[55] Shivani Upadhyay、Ronak Pradeep、Nandan Thakur、Daniel Campos、Nick Craswell、Ian Soboroff 和 Jimmy Lin。2025年。使用大语言模型和 UMBRELA 对相关性评估进行的大规模研究。载于*ICTIR*。358–368。[56] Shivani Upadhyay、Ronak Pradeep、Nandan Thakur、Nick Craswell 和 Jimmy Lin。2024年。UMBRELA：UMbrela 是必应相关性评估器的（开源复现版本）。*arXiv* 预印本 *arXiv:2406.06519*（2024年）。[57] Orion Weller、Kathryn Ricci、Eugene Yang、Andrew Yates、Dawn Lawrie 和 Benjamin Van Durme。2025年。Rank1：用于信息检索重排序的测试时计算。载于 *COLM*。[58] Eugene Yang、Andrew Yates、Kathryn Ricci、Orion Weller、Vivek Chari、Benjamin Van Durme 和 Dawn Lawrie。2025年。Rank-K：用于列表式重排序的测试时推理。*arXiv* 预印本 *arXiv:2505.14432*（2025年）。