仍然新鲜吗？评估检索基准中的时间漂移

Nathan Kuissi*
ngkuissi@uwaterloo.ca 滑铁卢大学 滑铁卢，加拿大

Suraj Subrahmanyan*
s2subrah@uwaterloo.ca 滑铁卢大学 滑铁卢，加拿大

Nandan Thakur†
nandan.thakur@uwaterloo.ca 滑铁卢大学 滑铁卢，加拿大

Jimmy Lin
jimmylin@uwaterloo.ca 滑铁卢大学 滑铁卢，加拿大

arXiv:2603.04532v1 [cs.IR] 4 Mar 2026

摘要

信息检索（IR）基准测试通常遵循克兰菲尔德范式，依赖静态且预先定义的语料库。然而，技术语料库中的时序变化，例如 API 弃用和代码重组，可能会使现有基准测试过时。在本研究中，我们探讨了时序语料漂移如何影响 FreshStack——一个专注于技术领域的检索基准测试。我们考察了 FreshStack 在 2024年10月 和 2025年10月 的两个独立语料库快照，以回答有关 LangChain 的问题。分析显示，2024 年提出的查询中除一个以外，其余查询均仍得到 2025年 语料库的完全支持，因为相关文档会从 LangChain“迁移”到竞争对手的代码仓库，例如 LlamaIndex。接下来，我们比较了检索模型在两个快照上的准确率，发现模型排名仅发生轻微变化；在 Recall@50 上，整体相关性最高达到 0.978 的 Kendall τ。这些结果表明，使用不断演变的时序语料库重新评估的检索基准测试，仍可可靠地用于检索评估。我们在 https://github.com/fresh-stack/driftbench 上公开发布了所有相关产物。

CCS 概念

• **信息系统** → **检索结果评估**；测试集合。

关键词

T时间分析；检索基准；评估 on

1 引言

信息检索（IR）基准测试依赖静态测试集合来评估和比较检索系统 [45, 35]。在克兰菲尔德范式下，测试集合通过固定的文档语料库、一组查询和相关性判断构建而成。诸如 MS MARCO [24], 、TREC 深度学习（DL） [10, 8, 9, 11], 和 BEIR [32] 等热门信息检索基准测试均采用这一方法，在基准测试构建期间创建固定的语料库快照和相关性判断。

测试集合有助于建立可复用的基线 [38,5, 36], ，但信息会发生变化，语料库也会持续演进并产生时间变化 [37, 15, 16, 26]。现实世界的文档集合（例如技术文档）具有高度动态性，因为代码经常被创建、删除和重组。随着时间推移，基准测试中捕获的文档集合快照会逐渐偏离该领域的当前状态。

在我们的工作中，我们评估了**FreshStack**中的时间漂移，该工作由Thakur等人 [31],利用LangChain技术文档的时间快照提出。我们使用十个为检索增强生成（RAG）提供文档的GitHub仓库的两个独立语料库快照来构建测试集合

*共同贡献。† 联系邮箱：nandan.thakur@uwaterloo.ca

框架与开发者工具，通过在2024年10月复现并将其扩展至2025年10月。在此期间，LangChain的文档因重组与弃用减少了67%，而Chroma则随着文档迁移增长了2.6× 。

我们复现了语料库构建过程，并在两个快照中分别生成相关性判断。我们采用混合融合方法检索可能相关的文档，将 BM 25 [29], BGE（Gemma-2） [6], E5 Mistral (7B) [40],和 Qwen3 (4B) 嵌入模型 [44]结合起来。接下来，我们使用 Cohere 的 Command A 作为评审者 [7]，进行要点级支持评估 [27, 1] ，从而自动构建两个时间快照下的测试集合，涵盖 LangChain 中的 203 个查询 [31]。我们的工作探讨以下三个研究问题：

研究问题1 现有查询能否在随时间动态变化的语料库中找到依据?
研究问题2 不同代码仓库中的相关文档分布会如何随时间变化?
研究问题3 面对技术文档中的时间漂移，模型排名是否仍保持一致?

据我们所知，本研究首次评估了高度动态语料库中的时间变化，具体而言，是小众领域的技术文档。以往研究评估过更通用语料库中的时间变化，例如新闻、 [37],联邦文件 [30], [4, 或 13] LongEval 中的科学文献集。本研究的主要贡献在于评估多个检索系统在动态变化语料库中的表现，从而能够衡量其性能随时间的变化。

实验结果表明，令人意外的是，除一个查询外，所有查询在两个语料库快照中都仍有充分依据且能够得到答案，尽管 LangChain 等代码仓库经历了大幅重组。这一现象的主要原因在于，2024年快照中 LangChain 里的相关文档，在2025年快照中分布于相关的 GitHub 仓库（如 LlamaIndex）。最后，文档检索任务的模型排名仍高度相关（Kendall τ = 0.978，按 Recall@50 计算），这表明尽管语料库受到时间变化的影响，基准测试仍能稳健地评估检索系统。

2 背景与相关工作

时间漂移是指语料库、查询及其相关性格局随时间发生的系统性变化，其成因包括文档的添加、删除、更新或重新组织。

近期关于评估测试集合中时间漂移的研究指出，基准测试的有效性可能会随时间推移而降低 [26], 这些研究主要关注判断变化或查询漂移，而非语料库漂移。当语料库发生大规模重组时，基准测试会如何表现，仍不清楚。其他研究与我们的实验在概念上有相似之处 [13, 15]。不过，这些探索聚焦于网页搜索数据，其中的查询

表 1：LangChain GitHub 仓库统计数据及 2024年和 2025年快照中的支持文档分布

GitHub 仓库	2024年10月		2025年10月	
	#文档	支持 (%)	#文档	支持 (%)
langchain	11,037	2,000 (50.9%)	3,628	921 (24.8%)
LangChainJS	3,852	733 (18.6%)	3,921	950 (25.5%)
llama_index	11,627	634 (16.1%)	9,751	842 (22.6%)
transformers	13,580	173 (4.4%)	13,798	223 (6.0%)
OpenAI 食谱	537	160 (4.1%)	970	320（8.6%）
chroma	1,359	127（3.2%）	2,949	245（6.6%）
Azure OpenAI 示例	413	49 (1.2%)	1,717	125 (3.4%)
openai-python	654	20 (0.5%)	1,274	25（0.7%）
azure-search-openai-demo	406	19（0.5%）	1,052	36（1.0%）
LangChain-nextjs-template	93	16 (0.4%)	105	32 (0.9%)
总计	**43,558**	**3,931**	**39,165**	**3,719**

简短而精炼。相比之下，我们的分析聚焦于更复杂的问题场景，这对受测检索系统构成了严峻挑战。此外，Cancellieri 等人 [4] 通过改变不同时期的语料库和查询，探讨了类似主题。不过，在这项工作中，我们将研究范围限定于技术文档这类动态语料库。在编程场景中，查询通常与同一任务相关，但由于文档发生变化，实现可能因版本而异。因此，在我们的研究背景下，动态语料库足以提供考察材料。最后，Soboroff [30] 回顾了时间变化对数据集本身的影响，而我们考察的是检索系统性能，而不只是基准测试。

3 实验设置与细节

评估检索系统需要由问答对构建的判定池。此外，我们需要一个足够大的语料库，以提供支撑问题答案所需的相关上下文。在先前关于 Fresh- Stack 的工作基础上，[31], 我们使用最初于2024年10月构建的 Stack Overflow 查询，并于2025年10月扩展该测试集。[1]

此外，我们将探索重点放在 **LangChain** 领域。虽然观察多个领域有助于全面分析时间变化对检索性能的影响，但聚焦于单一领域能让我们更有效地研究时间变化。LangChain 是开展时间探索的代表性领域，因为其 GitHub 仓库是自然语言处理领域中最活跃、变化最快的仓库之一 [21]。

我们的流程遵循 FreshStack [31]，并按照以下步骤，在任意时间点自动创建检索基准测试。以下小节将对每个步骤进行说明。

(1) **语料库准备**：我们收集并切分两个版本的 LangChain 相关代码仓库，以准备语料库。(2) **要点生成**：我们利用 Stack Overflow 上的问题和答案生成要点，即关键事实。(3) **预言机检索**：我们从语料库中检索一部分可能与回答查询相关的文档块。(4) **信息要点级评估**：对于每个查询，我们利用这部分文档块判断每个文档块是否支持某个要点（即回答问题所需的关键事实）。

[1]https://huggingface.co/datasets/freshstack/queries-oct-2024

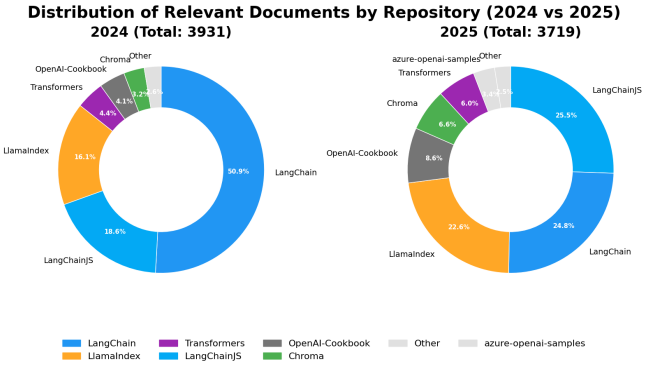

图 1：按 GitHub 仓库划分的 2024年和 2025年相关文档分布（%）示意图。

3.1 语料库准备

这里的目标是从两个不同时间快照中的 LangChain 相关代码仓库版本收集并构建语料库。此外，我们将多个代码仓库合并成一个语料库，因为 Stack Overflow 上标记为 LangChain 的查询可能需要其他代码仓库中的上下文。例如，某个 LangChain 问题可能涉及如何正确加载 HuggingFace 变换器模型，或如何使用 ChromaDB。表 1 列出了用于构建语料库的十个相关代码仓库。

为实现流程自动化，我们使用截至 2024年10月和 2025年10月各 GitHub 仓库主分支的快照。给出仓库名称后，我们会搜索这些日期之前的 GitHub 提交，并采用最新提交的哈希值。然后，我们将所有文档、代码文件、Jupyter 笔记本等进行 **分块**，每块最多包含 2048 个词元。为创建唯一的文档标识符，我们使用仓库名称、本地仓库中的相对文件路径，以及用于消歧的分块起始和结束字节信息。

3.2 要点生成

要点是构建答案所必需的原子级关键事实 [19, 20, 27, 31]。我们从 Stack Overflow 的查询和答案中自动生成要点，以拆解长篇答案并便于进行文档级评估。这些要点在预言机检索阶段也至关重要地充当查询 [31]。将要点用作评估工具的做法在检索系统性能领域日益增多 [1, 27]。

同样，大语言模型展现出从文档中提取相关且简洁信息的强大能力，可与繁琐且耗时的人工标注过程相媲美 [27, 23]。因此，对于 LangChain 中的每个查询，我们复用了在 FreshStack 中使用 GPT-4o 生成的要点 [31]。

3.3 预言机检索

要创建测试集，我们需要判定池。换言之，我们必须检索并标注与每个查询相关的文档。先前的研究 [43, 25, 31] 表明，在构建判定池时使用多样化的检索模型十分重要。

检索技术。 首先，我们使用 BM25 [29],，它能够检索在词汇层面相似的文档。接下来使用的模型是 BGE（Gemma-2） [6],

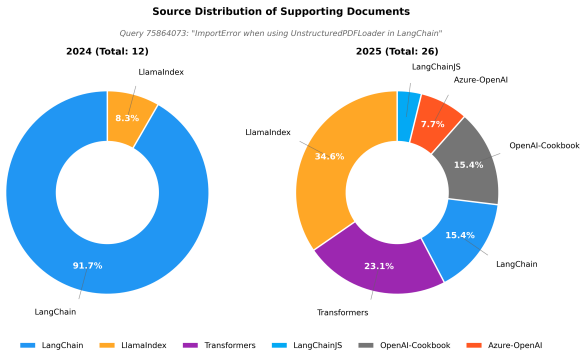

图2：LangChain查询 75864073在2024年与2025年语料库快照之间的来源分布变化。

一种稠密检索器模型[2]，嵌入维度为 3584，上下文长度为 8K。接下来，E5 Mistral (7B) [40] 是另一种稠密检索器模型[3]，嵌入维度为 4096。此外，我们使用 Qwen3 (4B)[4]，其嵌入维度为 4096，上下文长度为 32K。Qwen3嵌入在多个领域的检索任务中达到了最先进的准确率 [44]。最后，我们将每个模型的检索分数归一化到 $[0, 1]$ 与……之间，并通过混合融合将它们相加。

检索设置。 为了进一步提高判定池的多样性，我们在预言机设置中采用不同方法来构建查询。我们首先使用Stack Overflow上提供的原始答案。接下来，我们使用GPT-4o生成的要点，其中包含简洁且结构化的信息。然后，我们使用Qwen3-4B-Instruct模型生成子问题。[5] 我们不使用原始冗长的用户查询，而是使用Qwen3将其分解为子问题，从而通过思维链提示实现有效检索 [41]。此外，我们使用Qwen3为该Stack Overflow查询生成闭卷答案。最后，针对上述每种技术，我们使用融合方法收集排名前50的文档，并将其传递到下一阶段。

3.4 信息要点级评估

判断一篇文档是否与给定查询相关，传统上依赖于人工评审员。然而，这一过程在成本和时间方面都可能十分昂贵。近期研究表明，利用大语言模型作为相关性判断的自动评审员具有潜力 [33, 28, 12, 34]。信息检索测试集合通常包含较短的查询，这使得大语言模型更容易完成判断任务。由于Stack Overflow问题中的查询较为冗长，我们选择改为依据支持要点的文档来判断相关性。

我们使用**Cohere 的 Command A**[6] 作为评判器。该模型拥有 111B 个参数和 256k 的上下文长度，其性能已展现出可与 GPT-4o [7]媲美的水平。我们向 Command A 提供查询和答案（作为上下文）、生成的要点以及一组检索到的文档。最后，我们要求它识别哪些文档支持

[2]BGE (Gemma-2 2 3): https://huggingface.co/BAAI/bge-multilingual-gemma E5 Mistral 7B: https://huggingface.co/intfloat/e5-mistral-7b-instruct [4]Qwen3 (4B) 嵌入模型: https://huggingface.co/Qwen/Qwen3-Embedding-4B [5] Qwen3-4B-Instruct: https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507
[6]Cohere Command A：https://cohere.com/blog/command-a

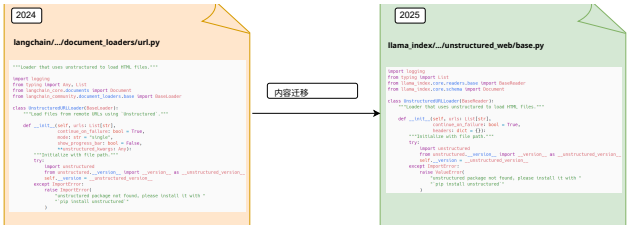

图3：UnstructuredURLLoader 类针对 LangChain 查询 75864073 从 LangChain（2024年）迁移，并集成到 LlamaIndex（2025年）。

要点中包含的信息。换句话说，对于给定查询，当且仅当文档支持至少一个要点的信息时，我们才将该文档定义为相关文档。

4 实验结果与分析

我们的分析主要聚焦于比较所选仓库两个快照中的语料库及各种模型的表现。首先，我们探讨 2023 至 2024 年提出的 LangChain 问题，是否仍能通过最新的 2025年语料库得到解答（第 4.1 节）。随后，我们进一步分析 GitHub 仓库中相关文档方面的一些变化（第 4.2 节）。分析这些变化有助于我们更好地理解技术文档检索任务的复杂性。最后，我们评估不同检索系统在两个时间快照中的表现（第 4.3 节）。

4.1 查询的时间支持

本节探讨**RQ1**：随着时间推移，现有查询是否仍以文档语料库为依据。例如，“代理”曾是 2024年 LangChain 仓库的一部分。然而，自 2025年起，智能体系统的实现已转移至 langGraph [39]。因此，关于智能体系统的查询将无法 从 2025年 LangChain GitHub 仓库中检索到相关文档。

实验设置。 为确定 LangChain 语料库的动态变化是否会限制先前提出的查询集在时间上的适用性，我们进行了以下实验：如果某个查询的所有要点都至少得到语料库中一篇相关文档的支持，那么语料库中就有足够的信息，可为构建 Stack Overflow 答案所需的关键事实提供依据。

结果。 我们在语料库的两个快照上进行了实验，发现只有 1 个要点未得到 2025年语料库的支持。也就是说，在 640 个要点（涵盖 203 个查询）中，除一个之外，2025年语料库中都至少有一篇相关文档支持其所提供的信息。因此，我们可以得出结论：即使存在时间变化，2025年和 2024年语料库快照都能为现有查询提供依据，并提供回答 LangChain 中问题所需的相关文档。

4.2 语料库的时间变化

在本节中，我们通过分析两个快照中相关文档的差异，探讨**研究问题 2**。此前，我们曾提出一种可能的情况：实现细节（例如代理）

表 2：使用2024年和2025年时间快照在 LangChain 上进行文档检索。评估所用指标为 α-nDCG@10、覆盖率@20和Recall@50。

嵌入模型	LangChain（2024年）			LangChain（2025年）		
	αN@10	C@20	R@50	αN@10	C@20	R@50
BM25 [29]	0.228	0.576	0.170	0.187	0.489	0.154
Qwen3（8B） [44]	**0.503**	**0.877**	0.432	0.480	0.868	0.424
Qwen3 (4B) [44]	0.501	0.868	**0.436**	**0.482**	**0.873**	**0.437**
Qwen3 (0.6B) [44]	0.442	0.821	0.362	0.411	0.782	0.372
Stella（1.5B） [42]	0.444	0.824	0.374	0.430	0.830	0.371
Stella (400M) [42]	0.386	0.812	0.340	0.388	0.805	0.345
Granite-Emb.-R2 [3]	0.396	0.767	0.318	0.359	0.766	0.325
Granite-Emb.-R2（小型） [3]	0.393	0.768	0.311	0.375	0.752	0.311
GTE（ModernBERT-base） [17]	0.365	0.716	0.271	0.343	0.728	0.279
GTE（大型）v1.5 [17]	0.324	0.709	0.239	0.276	0.683	0.242
E5 Mistral (7B) [40]	0.464	0.851	0.395	0.434	0.828	0.388
BGE（Gemma-2） [6]	0.423	0.794	0.341	0.401	0.828	0.348
Jina v4 [14]	0.408	0.811	0.381	0.417	0.821	0.379
Voyage-4-nano [2]	0.441	0.811	0.399	0.426	0.824	0.396

会从一个仓库迁移到另一个仓库。因此，考虑语料库中所选代码仓库内发生此类变化的影响十分重要。

实验设置。 我们考察以下问题：在被 Command A 判定为相关的文档中，它们与哪些 GitHub 仓库相关联？我们在表 2 和图 1 中展示支持文档的来源分布。

结果。 图 1 凸显了两个快照之间文档相关性的潜在分布变化。首先，我们观察到源自 LangChain 仓库的相关文档有所减少。2024年，LangChain 是全部相关文档中 50.9% 的来源。然而，到 2025年，这一比例降至 24.8%（略低于 LangChainJS 的 25.5%）。此外，在 2025年，我们观察到相关文档的来源更加多样化。这一变化与先前工作 [22], 一致，该工作将 LangChain 描述为一个依赖 LlamaIndex 等相关代码仓库的模块化生态系统。

案例研究。 作为一项案例研究，我们考察查询 75864073（标题: *LangChain* 中 *UnstructuredPDFLoader* 的导入错误），如图 3 所示。首先，我们观察到相关文档从 12 篇增加到 26 篇，这表明相关框架在功能上日益重叠，从而产生冗余，而不是转而依赖原始来源。2024年，LangChain 是相关文档的主要来源（91.7%）。然而，到 2025年，相关文档分布在六个代码仓库中，其中 LlamaIndex（34.6%）是最大来源。如图 3 所示，UnstructuredURLLoader 类（查询 75864073 的一篇相关文档）已从 LangChain 迁移到 LlamaIndex。

我们推测这一现象有两个原因：（1）LangChain 在 2025年重新组织了文档，并删除了有关第三方集成的内容；（2）LlamaIndex 和 Transformers 等竞争框架在 2025年扩展了文档，以涵盖类似的处理工作流。

这些观察再次凸显了技术文档检索任务固有的复杂性。检索系统不能只是在两个语料库快照中查找完全相同的文件

表 3：LangChain 2024 年和 2025 年语料库快照中各项指标对应模型排名的 Kendall τ 相关系数。

指标	α-nDCG@10	覆盖率@20	Recall@50
Kendall τ	0.846	0.692	0.978

来检索相关信息。由于仓库中发生了时间变化，稳健的检索系统必须关注运行时呈现的底层文档结构和内容，尽管缺乏时间一致性。

4.3 基准分析

在本节中，我们尝试回答**RQ3**。我们在两个快照上对多种嵌入模型的检索准确率进行基准测试，并加以比较，以观察模型排名是否发生变化。

实验设置与结果。除第 3.3 节中使用的检索模型外，我们还对以下系列的模型进行基准测试：Qwen3嵌入 [44],Stella [42], Granite-R2 [3],GTE [18], Jina V4 [14] 以及 Voyage-4-nano [2]。在评估中，我们遵循 FreshStack [31] 中的设置，并使用 α-nDCG@10、覆盖率@20 和 Recall@50 来衡量检索多样性与相关性。各项指标的定义见 Thakur 等人 [31]。文档检索结果见表 2。总体而言，我们观察到 Qwen3 (4B) 与 (8B) 取得了最高分数。对于 α-nDCG@10 和 Recall@50，我们总体上观察到，在 2024年与 2025年快照之间，大多数检索模型的分数有所下降。

模型排名相关性。 我们使用 Kendall τ 相关性指标来衡量 2024年和 2025年语料库快照之间模型排名的变化。从表 3 可以看出，两个快照中的模型排名呈现出很强的相关性，尤其是 Recall@50（Kendall $\tau = 0.978$）和 α-nDCG@10（Kendall $\tau = 0.846$）。不过，覆盖率@20 的相关性相对较弱（Kendall $\tau = 0.692$）。

这种正相关表明，尽管语料库不断变化，不同时期快照中的模型在检索**相关信息**方面的能力仍相当稳定。然而，在检索多样化的相关段落方面，这些模型似乎表现不佳，而要综合出全面的 Stack Overflow 答案，就需要这些段落涵盖所有要点。

5 结论与未来工作

在这项工作中，我们研究了语料库随时间发生的变化如何影响 LangChain 子集上的 Fresh-Stack。我们首先发现，在 2025年10月的语料库快照中，除一个查询外（202/203），其余查询的每个要点都有受支持文档（即相关文档）。接下来，我们探究这类查询效用较高的原因。我们发现，函数与类经常迁移到其他竞争框架，例如从 LangChain 迁移到 →LlamaIndex。最后，我们发现，语料库两个时间快照中的模型排名高度相关，这表明尽管语料库随时间发生变化，检索系统在检索相关文档方面依然保持一致。这些发现表明，Fresh-Stack（LangChain）能够适应技术代码文档的动态变化。

我们工作的一个关键方面，是将多个 GitHub 仓库构建的测试集作为语料库。由于 LangChain 的结构特点，即使某个仓库中的代码功能已被弃用，其中许多功能仍会保留，或迁移至相关代码库或竞争对手的代码库。不过，也有必要考虑其他领域的情况是否类似。例如，在维基百科中，语料库中的信息可能会随时间变化，甚至问题的答案也可能发生演变。在这类情境下，可能需要在构建测试集之前，为每个时间快照生成要点。此外，研究检索系统在不同时间快照中的性能能否保持一致，也会很有意思。

致谢

我们感谢 Cohere 为我们的实验提供 API 额度，以便运行 Command A 作为评判模型。本研究部分得到了加拿大自然科学与工程研究委员会（NSERC）的支持。

参考文献

[1] Zahra Abbasiantaeb、Simon Lupart、Leif Azzopardi、Jeffrey Dalton 和 Mohammad Aliannejadi。2025年。对话式金标准：使用黄金片段评估个性化对话式搜索系统。载于 第 *48* 届国际 *ACM SIGIR* 信息检索研究与发展会议论文集 （SIGIR ’25）。计算机协会，意大利帕多瓦，3455–3465。isbn：9798400715921。doi:10.1145/3726302.3730316。[2] Voyage AI。2026年。Voyage 4 模型系列：采用混合专家架构的共享嵌入空间。https://blog.voyageai.com/2024/09/18/voyage-3/。[3] Parul Awasthy 等人。2025年。Granite embedding R2 模型。*CoRR*，abs/2508.21085。[4] Matteo Cancellieri 等人。2025年。CLEF 2025 的 LongEval：信息检索模型性能的纵向评估。载于 信息检索进展：第 *47* 届欧洲信息检索会议，*ECIR 2025*，意大利卢卡，*2025*年*4*月*6–10*日，论文集，第 *V* 部分。施普林格出版社，意大利卢卡，382–388。isbn：978-3-031-88719-2。doi:10.1007/978-3-031-88720-8 58。[5] Ben Carterette、Evgeniy Gabrilovich、Vanja Josifovski 和 Donald Metzler。2010年。测量测试集合的可复用性。载于 第三届 *ACM* 国际网络搜索与数据挖掘会议论文集 （WSDM ’10）。计算机协会，美国纽约州纽约，231–240。isbn：9781605588896。doi:10.1145/1718487.1718516。[6] Jianlyu Chen、Shitao Xiao、Peitian Zhang、Kun Luo、Defu Lian 和 Zheng Liu。2024年。M3-embedding：通过自知识蒸馏实现多语言性、多功能性和多粒度文本嵌入。载于 计算语言学协会论文集：*ACL 2024*。Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，（2024年8月），2318–2335。doi:10.18653/v1/2024.findings-acl.137。[7] Cohere。2025年。Command A：面向企业的大语言模型。*CoRR*，abs/2504.00698。[8] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Campos 和 Jimmy Lin。2022年。TREC 2021 深度学习赛道概览。载于 文本检索会议（*TREC*）。美国国家标准与技术研究院。TREC，（2022年5月）。https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2021-deep-learning-track/。[9] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Daniel Campos、Jimmy Lin、Ellen M. Voorhees 和 Ian Soboroff。2023年。TREC 2022 深度学习赛道概览。载于 文本检索会议（*TREC*）。美国国家标准与技术研究院。TREC，（2023年3月）。https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2022-deep-learning-track/。[10] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、DanielCampos、EllenM. Voorhees 和 Ian Soboroff。2021年。TREC 深度学习赛道：大数据规模下的可复用测试集合。载于 第 *44* 届国际 *ACM SIGIR* 信息检索研究与发展会议论文集 （SIGIR ’21）。计算机协会，虚拟活动，加拿大，2369–2375。isbn：9781450380379。doi:10.1145/3404835.3463249。[11] Nick Craswell、Bhaskar Mitra、Emine Yilmaz、Hossein A. Rahmani、Daniel Campos、Jimmy Lin、Ellen M. Voorhees 和 Ian Soboroff。2024年。TREC 2023 深度学习赛道概览。载于 文本检索会议（*TREC*）。美国国家标准与技术研究院。TREC，（2024年2月）。https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2023-deep-learning-track/。[12] Naghmeh Farzi 和 Laura Dietz。2025年。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论文集（*ICTIR*）（ICTIR ’25）。计算机协会，意大利帕多瓦，254–263。isbn：9798400718618。doi:10.1145/3731120.3744591。[13] Petra Galuscáková、Romain Deveaud、Gabriela González Sáez、Philippe Mulhem、Lorraine Goeuriot、Florina Piroi 和 Martin Popel。2023年。LongEval-Retrieval：用于连续网络搜索评测的法英动态测试集合。载于国际 *ACM SIGIR* 信息检索研究与发展会议第 *46* 届会议论文集（SIGIR ’23）。计算机协会，台北，台湾，3086 3 539618 3094 9781450394086 10 1145 3591921 – 。isbn：。doi：。/。。[14] Michael Günther 等。2025 年。jina-embeddings-v4：用于多模态多语言检索的通用嵌入。*CoRR*，abs/2506.18902。[15] Jüri Keller、Timo Breuer 和 Philipp Schaer。2024 年。信息检索测试集合中的时间变化评估。载于*2024* 年 *ACM SIGIR* 信息检索理论国际会议论文集（ICTIR ’24）。计算机协会，华盛顿哥伦比亚特区，美国，3–13。isbn：9798400706813。doi:10.1145/3664190.3672530。[16] Jüri Keller、Timo Breuer 和 Philipp Schaer。2024 年。纵向信息检索评测的可复现性度量。载于实验性信息检索与多语言、多模态和交互：*CLEF* 协会第 *15* 届国际会议，*CLEF 2024*，法国格勒诺布尔，*2024* 年 *9* 月 *9–12* 日，论文集，第一部分。施普林格出版社，法国格勒诺布尔，215–226。isbn：978-3- 031-71735-2。doi:10.1007/978-3-031-71736-9_16。[17] Zehan Li、Xin Zhang、Yanzhao Zhang、Dingkun Long、Pengjun Xie 和 Meishan Zhang。2023 年。通过多阶段对比学习迈向通用文本嵌入。*CoRR*，abs/2308.03281。[18] Zehan Li、Xin Zhang、Yanzhao Zhang、Dingkun Long、Pengjun Xie 和 Meishan Zhang。2023 年。通过多阶段对比学习迈向通用文本嵌入。*CoRR*，abs/2308.03281。[19] Jimmy Lin 和 Dina Demner-Fushman。2005 年。自动评估定义类问题的答案。载于人类语言技术会议暨自然语言处理经验方法会议论文集。Raymond Mooney、Chris Brew、Lee-Feng Chien 和 Katrin Kirchhoff（编）。计算语言学协会，加拿大温哥华、不列颠哥伦比亚，（2005 年 10 月），931–938。https://aclanthology.org/H05-1117。[20] Jimmy Lin 和 Dina Demner-Fushman。2006 年。自动评估复杂问题答案的方法。信息检索，9，5，（2006 年 11 月），565–587。doi:10.1007/s10791-006-9003-7。[21] Daniel Liu、Krishna Upadhyay、Vinaik Chhetri、A. B. Siddique 和 Umar Farooq。2026 年。关于多智能体人工智能系统发展及其问题的大规模研究。*CoRR*，abs/2601.07136。[22] Rajesh Kumar Malviya、Vishal Javalkar 和 Rano Malviya。2024 年。LangChain、LlamaIndex 和 Haystack 在企业级人工智能客户支持系统中的可扩展性与性能基准测试。载于*IJGIS 2024* 年秋季会议。https://ijgis.pubpub.org/pub/6yecqicl。新世界基金会。（2024 年 9 月）。[23] Dasha Metropolitansky 和 Jonathan Larson。2025 年。迈向事实性主张的有效提取与评估。载于计算语言学协会第 *63* 届年会论文集（第一卷：长篇论文）。Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编）。计算语言学协会，奥地利维也纳，（2025 年 7 月），6996–7045。isbn：979-8-89176-251-0。doi:10.18653/v1/2025.acl-long.348。[24] Tri Nguyen、Mir Rosenberg、Xia Song、Jianfeng Gao、Saurabh Tiwary、Rangan Majumder 和 Li Deng。2016 年。MS MARCO：人工生成的机器阅读理解数据集。载于认知计算研讨会：神经与符号方法的整合 *2016*（与神经信息处理系统第 *30* 届年会（*NIPS 2016*）同期举行），西班牙巴塞罗那，*2016* 年 *12* 月 *9* 日（CEUR 研讨会论文集）。Tarek Richard Besold、Antoine Bordes、Artur S. d’Avila Garcez 和 Greg Wayne（编）。第 1773 卷。CEUR-WS.org。https://ceur-ws.org/Vol-1773/CoCoNIPS%5 C_2016%5C_paper9.pdf。[25] Ioannis Papadimitriou、Ilias Gialampoukidis、Stefanos Vrochidis 和 Ioannis Kompatsiaris。2024 年。RAG playground：用于系统评估 RAG 系统中检索策略和提示工程的框架。*CoRR*，abs/2412.12322。[26] Andrew Parry、Maik Fröbe、Harrisen Scells、Ferdinand Schlatt、Guglielmo Faggioli、Saber Zerhoudi、Sean MacAvaney 和 Eugene Yang。2025 年。相关性判断的变化与测试集合的有效期。载于国际 *ACM SIGIR* 信息检索研究与发展会议第 *48* 届会议论文集（SIGIR ’25）。计算机协会，意大利帕多瓦，3387–3397。isbn：9798400715921。doi:10.1145/3726302.3730308。[27] Ronak Pradeep、Nandan Thakur、Shivani Upadhyay、Daniel Campos、Nick Craswell、Ian Soboroff、Hoa Trang Dang 和 Jimmy Lin。2025 年。The Great Nugget Recall：使用大语言模型自动进行事实提取和 RAG 评测。载于国际 *ACM SIGIR* 信息检索研究与发展会议第 *48* 届会议论文集（SIGIR ’25）。计算机协会，意大利帕多瓦，180–190。isbn：9798400715921。doi:10.1145/3726302.3730090。[28] *LLMJudge*：用于相关性判断的大语言模型，（2024 年 8 月）。https://www.microsoft .com/en-us/research/publication/llmjudge-llms-for-relevance-judgments/。

[29] Stephen Robertson 和 Hugo Zaragoza。2009 年。概率相关性框架：BM25 3、4、333 389 及其他。信息检索领域的基础与趋势，–。[30] Ian Soboroff。2006 年。动态测试集合：衡量实时网络上的搜索效果。载于国际 *ACM SIGIR* 信息检索研究与发展会议第 *29* 届年会论文集（SIGIR '06）。计算机协会，美国华盛顿州西雅图，276–283。isbn：1595933697。doi:10.1145/1148170.1148220。[31] Nandan Thakur、Jimmy Lin、Sam Havens、Michael Carbin、Omar Khattab 和 Andrew Drozdov。2025 年。FreshStack：构建用于评估技术文档检索的真实基准测试。载于神经信息处理系统第三十九届年会数据集与基准测试专题。https://openreview.net/forum?id=54TTgXlS2U。[32] Nandan Thakur、Nils Reimers、Andreas Rücklé、Abhishek Srivastava 和 Iryna Gurevych。2021 年。BEIR：用于零样本评估信息检索模型的异构基准测试。载于神经信息处理系统数据集与基准测试专题论文集 *1*，*NeurIPS* 数据集与基准测试 *2021*，*2021* 年 *12* 月，线上。Joaquin Vanschoren 和 Sai-Kit Yeung（编）。https://datasets-be nchmarks-proceedings.neurips.cc/paper/2021/hash/65b9e ea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html。[33] 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。载于国际 *ACM SIGIR* 信息检索研究与发展会议第 *47* 届会议论文集（SIGIR '24）。计算机协会，美国华盛顿哥伦比亚特区，1930 3 657707 1940 9798400704314 10 1145 3626772 –。isbn：。doi：。/。。[34] Shivani Upadhyay、Ronak Pradeep、Nandan Thakur、Daniel Campos、Nick Craswell、Ian Soboroff 和 Jimmy Lin。2025 年。使用 umbrela 和大语言模型开展相关性评估的大规模研究。载于*2025* 年国际 *ACM SIGIR* 信息检索创新概念与理论会议论文集（*ICTIR*）（ICTIR '25）。计算机协会，意大利帕多瓦，358–368。isbn：9798400718618。doi:10.1145/3731120.3744605。[35] Ellen Voorhees。2009 年。我不是来埋葬 Cranfield 的，而是来赞美它的。英文。载于 HCIR 2009：连接人机交互与信息检索，华盛顿哥伦比亚特区，-1，（2009-10-26 2009）。https://www.nist.gov/publications/i-co me-not-bury-cranfield-praise-it。[36] Ellen M. Voorhees。2018 年。使用多臂老虎机技术构建公平且可复用的测试集合。载于第 *27* 届 *ACM* 信息与知识管理国际会议论文集（CIKM '18）。计算机协会，意大利都灵，407 32692 06 416 9781450360142 10 1145 3271766 –。isbn：。doi：。/。。[37] Ellen M. Voorhees、Ian Soboroff 和 Jimmy Lin。2022年。旧的 TREC 测试集能否可靠地评估现代神经检索模型？*CoRR*，abs/2201.11086。[38] Ellen M. Voorhees 和 Dawn M. Tice。2000年。构建问答测试集合。收录于第*23*届年度国际 *ACM SIGIR* 信息检索研究与发展会议论文集 （SIGIR '00）。计算机协会，雅典，希腊，200–207。isbn: 1581132263。doi:10 .1145/345508.345577。[39] Jialin Wang 和 Zhihua Duan。2024年。智能 Spark 智能体：面向可扩展、可视化且增强的大数据机器学习工作流的模块化 lang-graph 框架。*CoRR*，abs/2412.01490。[40] Liang Wang、Nan Yang、Xiaolong Huang、Linjun Yang、Rangan Majumder 和 Furu Wei。2024年。利用大语言模型改进文本嵌入。收录于第*62*届计算语言学协会年会论文集（第*1*卷 *16*，*2024*年 *2024*年，*11*：长论文），*ACL* 曼谷，泰国，*8*月 -。Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）计算语言学协会，11897 64 2 11916 10 18653 1 2024年 –。doi: . /V / .ACL-LONG. 。[41] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。2022年。思维链提示激发大语言模型中的推理。收录于第*36*届神经信息处理系统国际会议论文集 （NIPS '22）文章编号 1800。Curran Associates Inc.，新奥尔良，路易斯安那州，美国，14页。isbn: 9781713871088。[42] Dun Zhang 和 FulongWang。2024年。Jasper 与 Stella：SOTA 嵌入模型的蒸馏。*CoRR*，abs/2412.19048。[43] Xinyu Zhang、Nandan Thakur、Odunayo Ogundepo、Ehsan Kamalloo、David Alfonso-Hermelo、Xiaoguang Li、Qun Liu、Mehdi Rezagholizadeh 和 Jimmy Lin。2023年。MIRACL：覆盖18种不同语言的多语言检索数据集。计算语言学协会会刊，11，（2023年9月），1114 1131 10 1162 _ _00 – 。595 2157340 00595 10 1162 00595 eprint: https://direct.mit.edu/tacl/article-pdf/doi/ . /tacl a / /tacl a .pdf。doi: . / tacl a 。[44] Yanzhao Zhang 等人。2025年。Qwen3 嵌入：借助基础模型推进文本嵌入与重排序。*CoRR*，abs/2506.05176。[45] Justin Zobel。1998年。大规模信息检索实验的结果有多可靠？收录于第*21*届年度国际 *ACM SIGIR* 信息检索研究与发展会议论文集 （SIGIR '98）。计算机协会，墨尔本，澳大利亚，307–314。isbn: 1581130155。doi:10.1145/290941.291014。