ArcBERT：用于探索整合多组学元数据的基于大语言模型的搜索引擎

Gajendra Doniparthi*、Shashank Balu Pandhare†、Stefan Deßloch* 和 Timo M¨uhlhaus‡
*异构信息系统，
†商用车辆技术中心，
‡计算系统生物学，凯撒斯劳滕-兰道莱茵兰-普法尔茨工业大学，戈特利布-戴姆勒街 47 号，67663 凯撒斯劳滕，德国 gajendra.doniparthi@cs.rptu.de，pandhare@rptu.de，stefan.dessloch@cs.rptu.de，timo.muehlhaus@rptu.de

arXiv:2512.15365v1 [cs.DB] 17 Dec 2025

摘要—**研究数据管理（RDM）生态系统中的传统搜索应用，对于帮助用户发现和探索研究数据集中的结构化元数据至关重要。通常，文本搜索引擎要求用户提交基于关键词的查询，而非使用自然语言。然而，使用基于特定领域内容训练的大语言模型（LLM）执行专门的自然语言处理（NLP）任务，正变得日益普遍。我们介绍 ArcBERT，这是一个旨在实现集成元数据探索的基于大语言模型的系统。与传统搜索应用不同，ArcBERT 能够理解自然语言查询，并依靠语义匹配。值得注意的是，ArcBERT 还能够理解元数据中的结构和层级结构，从而有效处理多种用户查询模式。**

关键词—**大语言模型、研究数据管理、注释研究上下文、组学元数据、文本搜索引擎**

I. 引言

生物科学研究界普遍认为，基于 FAIR 原则（即数据可发现、可访问、可互操作且可重用）的有效研究数据管理（RDM）平台，对于提高异构研究数据的可重复性、可见性和可访问性至关重要 [1]–[4]。搜索与探索应用在全面整合各个研究数据管理生态系统内部及彼此之间的标准化实验元数据方面发挥着关键作用。这些应用应能近乎实时地提供汇总的元数据视图，同时隐藏参与团体之间数据模型和数据交换方法的异构性。探索式查询处理技术使科学家能够交互式地探索元数据及其相关数据集、识别模式并生成数据驱动的洞见。MetaboLights [5], [6]和 PRIDE [7],等常用开源端点存储库，使用户能够直接进行自由文本元数据搜索。它们还提供搜索分面，让用户能够按特定研究或生物体筛选结果。

随着大语言模型（LLM）在生物科学研究领域日益普及，它们正成为探索标准化研究数据的传统文本搜索引擎的一种颇具前景的替代方案 [8]。现代语言模型能够理解搜索关键词及用户查询背后的语义，从而提供更准确的结果。大语言模型有望彻底改变我们分析整合实验元数据的方式，成为比传统文本搜索引擎更强大、更具洞察力的工具。

针对生物科学领域，已有一份在大规模生物语料库 [9] 上训练的大语言模型（LLM）综合清单。BioBERT [10]和 SciBERT [11]等领域专用大语言模型已在问答和文本挖掘等任务中展现出良好效果。这些模型首先在通用领域语料库（例如维基百科）上对 BERT [12] 模型进行预训练，随后再使用特定领域的语料库继续训练。BioBERT 使用来自 PubMed 和 PMC 等来源的生物医学文本进行微调，而 SciBERT 则使用来自 Semantic Scholar 的科学文献进行微调。针对特定的生物自然语言处理（NLP）任务对模型进行微调，能够显著提升其对生物学术语的理解能力。

使用大语言模型开发用于探索植物科学领域多组学元数据的系统时，需要考虑三个关键方面。首先，充分利用 RDM 解决方案及其功能，对于构建全面的整合组学元数据语料库至关重要。其次，需要对模型进行微调：使用基于植物科学文献的预训练模型，并通过组学元数据语料库对其进行调整，以增强模型以自然语言理解用户查询的能力。最后，由于元数据规范遵循层级结构，模型必须理解这些层级，才能有意义地匹配用户查询。本文的贡献如下。

- 我们重新审视了我们的研究数据管理方法（*PLANTdataHUB*）和现有的元数据搜索应用程序 *ARC* 元数据注册库，并介绍我们的研究数据管理解决方案如何构建并集成多组学元数据。
- 我们介绍了 *ArcBERT*，这是一种基于 Sentence-BERT [13] 的语言模型，用于集成元数据探索。该模型使用 PubMed [14] 中的植物科学文献进行预训练，并使用集成组学元数据进行微调。
- 我们提出使用索引机制为 ArcBERT 的向量嵌入建立索引，并借助 FAISS 库进行快速近邻搜索。
- 我们评估了使用 ArcBERT 进行

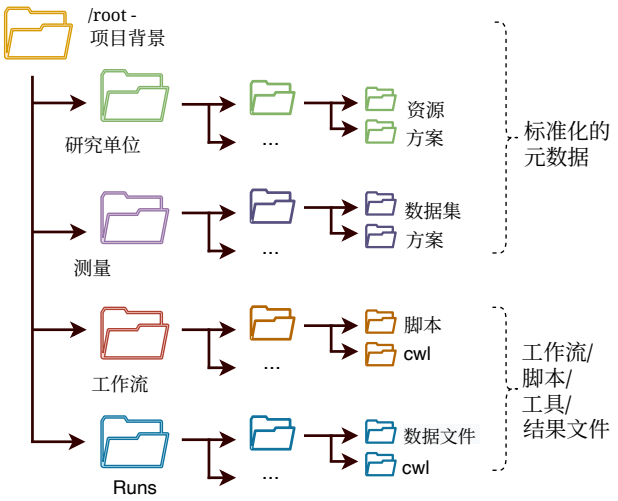

图 1. ARC 文件夹规范：将标准格式元数据与工作流、计算流程脚本以及工作流执行产生的结果文件/制品打包在一起。

元数据探索相较于传统文本搜索引擎的有效性。

以下章节将简要介绍我们的 PLANTdataHUB 生态系统及其核心组件。第 III 节概述 ARC 元数据注册库应用程序如何连接多个 DataHUB 并整合组学元数据。第 IV 节介绍 ArcBERT、使用注册表应用程序中的组学元数据进行预训练和微调，以及对 ARC 文档和结构嵌入进行索引。最后，在实验部分（第 V 节），我们将评估 ArcBERT 相较于 ElasticSearch 在元数据探索方面的效果。

II. PLANTDATAHUB

PLANTdataHUB 是一套全面的 FAIR 解决方案，旨在管理植物科学领域的研究数据 [15]。在这一研究数据管理解决方案中，研究产出单位称为 注释研究上下文（简称 ARC），它们是 FAIR 数字对象。ARC 是自包含、可互操作且可复现的，并具有唯一的数字对象标识符（DOI）。PLANTdataHUB 解决方案利用分布式版本控制系统（GitLab）存储并管理各个 ARC 的版本。它灵活支持维护实验室实验设计—测试—重复周期的每次迭代，从而避免了开发定制版本管理工具的需要。

A. 注释研究上下文

注释研究上下文（ARC）的核心理念是遵循固定的文件夹和文件布局，以通用且可互操作的格式组织实验元数据、测量数据、工作流、软件、外部文件等，从而确保研究项目中每项实验的可重复性 [16]。如图 1 所示，ARC 的格式经过标准化且具有灵活性，能够适应并支持任何组学研究中从简单到复杂的项目场景。ARC 规范旨在保持灵活性，能够容纳以任何已知元数据标准表示的实验元数据 [17]。

最初的植物研究数据 ARC 规范尤其支持 ISA 等知名标准。ISA 标准的抽象模型以分层的 ISA-JSON 文件格式实现 [18]。ARC 规范

还使用一种通用工作流语言来编写工作流和脚本，以构建计算流水线。工作流和计算分析产生的制品会存入各自的运行文件夹中。ARC 也可以编译为其他打包格式，例如 RO-Crate [19]。

B. DataHUB

在 PLANTdataHUB 模型中，集中式存储库称为 *DataHUB*。DataHUB 可以根据研究团队的数据共享需求和安全策略，部署在本地或远程。它们提供 ARC 的版本管理，并通过标准化的 API 访问，便于搜索和发现。此外，PLANTdataHUB 还提供各种辅助工具，便于在本地及远程 DataHUB 中整理、协作和维护 ARC。

III. 元数据集成

ARC 元数据注册库 是一款云原生应用，旨在为植物科学界提供实验元数据的集成搜索与分析。PLANTdataHUB 解决方案中的注册表应用程序旨在同时集成多个 DataHUB，并汇总呈现来自各个 DataHUB 中 ARC 顶层的数据实时视图。最重要的是，用户可以同时搜索和探索来自不同 ARC 以及不同 DataHUB 的数据，从而为集成元数据探索提供平台。ARC 元数据注册库提供一系列安全 API，用于访问标准化元数据；也就是说，当前版本支持分别以 JSON 和 JSON-LD 序列化的 ISA 和 RO-Crate 元数据（后者使用 Schema.org 注释）。

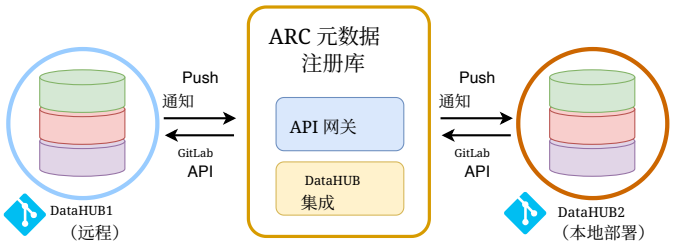

图 2. ARC 元数据注册库与多个本地部署及远程 DataHUB 集成，这些 DataHUB 托管着 ARC。

HUB 集成是一项至关重要的注册库应用服务，可集成多个 DataHUB，并即时下载 ARC 中的实验元数据。每个参与的 DataHUB 都需要进行一次性设置，以便与 ARC 元数据注册库建立双向数据交换，如图 2 所示。设置完成后，ARC 元数据注册库会实时接收推送

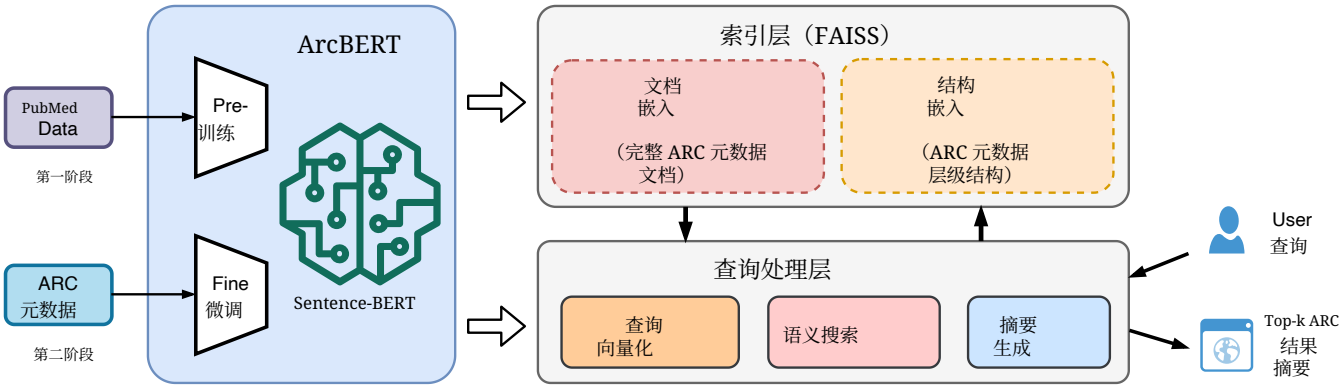

图3. ArcBERT架构，展示了Sentence-BERT模型、索引层和查询处理层。

通知：每当 GitLab 上的 ARC 存储库更新时，DataHUB 都会发送通知，而 HUB集成服务则会下载并处理最新变更。集成托管 ARC 的 DataHUB 有两个关键方面。第一，当用户更新 ARC 时，系统需要实时生成通知。例如，修改 ARC 中现有的工作流、修改管理元数据等。第二，通过 API 访问 DataHUB 中的 ARC，同时抽象通信方式。

Web用户界面服务维护供用户搜索和探索实验元数据的网页。

IV. 模型方法

大语言模型（LLM）基于Transformer架构构建，通常包含数亿至数十亿个可训练参数。它们在海量文本语料库上训练，因此擅长理解自然语言。这些模型还可以针对特定领域和学科进行微调。例如，来自Transformer的双向编码器表示（BERT） [12] 模型通常用于分类、命名实体识别（NER）和摘要生成等任务。相比之下，生成式预训练变换器（GPT） [20] 模型主要用于文本生成和翻译。

调整任何句子级语义模型的一个关键步骤，是将文本输入转换为能够捕捉语义含义的稠密向量表示。例如，Sentence-BERT [13], 会将每个句子转换为固定长度的向量，从而可以使用余弦距离测量语义相似度。即使表面措辞不同，表达相似观点的句子也会映射到这个向量空间中彼此接近的位置。使用整合多组学元数据进行微调后，该模型便能够理解自然语言形式的用户查询，并提供语义相似的查询结果；这与依赖关键词匹配的传统文本搜索引擎不同。

A. ArcBERT

我们提出了 ArcBERT，这是一种基于 Sentence-BERT 架构的领域专用语义搜索系统，如图 3 所示。该系统专为注释研究上下文（ARC）而设计，这些上下文包含来自各种组学研究的高度结构化元数据。为了根据语义相似度高效检索相关实验元数据，我们使用 FAISS 创建了一个稠密语义向量索引，从而能够针对任何给定的输入查询 [21]，快速检索语义上最相似的前 k 条 ARC 元数据。该模型和索引系统旨在保持语义相似度，并保留元数据中的层级关系，确保在不同元数据层之间实现有意义的查询匹配。

1) 预训练： 我们使用广泛应用的公共数字档案 PubMed 构建了一个特定领域的训练语料库，以训练模型，使其能够有效表征植物科学文献中的语言。我们选取了 2000 年 1 月至 2023 年 12 月期间发表、标题或摘要中包含“plant science”或“plant biology”关键词的英文文章。随后，我们对提取的数据集进行了清理，并整理为 JSON 对象列表，每个对象代表一篇文章。在训练 Sentence-BERT 模型之前，我们将每篇文章的相关字段（如标题、摘要、方法和结果）拼接成一个字符串。

2) 微调： 我们利用元数据中的层、特定字段注释和嵌套关系来改进模型，从而准确体现实际多组学数据的情况。ARC 规范提供了有关整体项目背景、主要贡献者、研究设计、分析类型和工作流的详细信息。通过使用按照 ARC 规范结构化的多组学元数据对模型进行微调，我们可以调整其嵌入，使其反映底层语义和上下文。我们直接利用 ARC 元数据注册库应用程序中的集成多组学数据快照。我们对数据集进行清理并展平为纯文本，如清单 1 所示，将标题、描述、研究、测定、贡献者以及任何可用的本体论参考等元素合并到“search text”字段中。

3）文档嵌入：对 ArcBERT 模型进行微调后，数据集中的每个 ARC 文档都会被展平为相应的“搜索文本”字段，随后转换为固定长度的嵌入向量。然而，随着嵌入数量增加，计算查询向量与模型

嵌入之间的相似度比较会变得十分耗费计算资源，导致暴力方法效率低下。我们使用 FAISS（一个专为稠密向量快速近邻搜索而设计的库）来加速查询处理。FAISS 提供多种索引结构；为此，我们使用 $IndexFlatIP$ 类型，该类型用于计算向量之间的内积。由于嵌入在预处理和微调过程中已归一化为单位长度，这些内积便对应于余弦相似度分数——这是句子级别语义比较中广泛使用的度量。当用户提交查询时，系统会通过 ArcBERT 模型处理该查询，从而在相同的向量空间中生成嵌入。然后将此嵌入与 FAISS索引进行比较，根据余弦相似度找出最相关的匹配项。

列表 1. 训练数据集片段，其中项目信息被展平为字符串“搜索文本”，用于微调模型


```
"title": "系统性研究对↪中度和急性高温的响应：␣绿藻↪莱茵衣藻。", "␣
description": "藻类培养物以↪ 混合营养方式（TAP培养基）生长。在经历
24小时的35/40摄氏度↪␣处理后，细胞被转回↪ 室温下48小时。随后采集
了组学样本。", "人员": ␣["Ningning Zhang（唐纳德·丹佛斯植物↪科学中心）
"], "出版物": "系统性分析揭示了绿藻莱茵衣藻对中度和急性高温的共同↪
及独特响应。", "研究": ␣["HeatstressExperiment"],"测定": ␣["蛋白质组学",
" 转录组学 ",↪" 代谢组学 ", " 生长 "],"␣搜索 _ 文本 ": " 系统性研究对 ↪ ␣中
度和急性高温的响应：绿藻↪ 莱茵衣藻。藻类培养物↪ 以混合营养方式生
长（TAP培养基）。在经历↪ 24小时的35/40摄氏度处理后，细胞被↪ 转
回室温下48小时。随后采集了组学↪ 样本。Ningning Zhang（唐纳德·↪
丹佛斯植物科学中心）系统性↪ 分析揭示了对↪中度和急性高温的共同及独
特响应，涉及绿藻↪莱茵衣藻。↪ HeatstressExperiment 蛋白质组学 转录
组学↪代谢组学 生长"
```


4) 结构嵌入：每个 ARC 文档的扁平化字符串都提供了全面的摘要。然而，在检索时将其作为单个数据块使用，会限制系统突出显示与用户查询相对应的特定元数据元素的能力。在科学数据分析中，用户通常关注特定层级的元数据，例如某项研究或测定名称、数据管理员或机构，以及与特定缓冲条件相关的实验参数值。

为增强搜索能力，ARC 元数据被拆分为更小、语义上更有意义的片段，称为“数据块”。每个数据块对应一个不同的元数据类别，例如研究、测定、参数、个人或出版物，并独立处理。这种方法使我们的系统既能匹配完整文档，也能精确定位 ARC层级结构中与用户查询相符的具体部分。例如，自然语言查询 莱茵衣藻的干旱胁迫实验可能与某项研究描述的关联更密切，而非标题或摘要。

微调阶段之后，数据集中的每个 ARC 文档都会被划分为数据块，并使用

表 I 语义索引的数据块数据结构格式。

字段	Type	描述
分块文本	字符串	从以下内容中提取的数据块 ARC 文档（例如，研究描述或测定名称）。
字段类型	字符串	ARC 结构中的字段名称。（例如，研究、测定、人员、出版物）
ARC 标识符	字符串	用于将数据块关联到其父级 ARC 文档的唯一标识符至其父级 ARC 文档。
数据块索引	整数	数据块在数据字段中的位置（用于在有多个条目时进行排序）。
嵌入	向量	由 ArcBERT 计算得到的稠密表示。

微调后的 ArcBERT 模型分别进行嵌入。随后，这些嵌入会存储在 FAISS索引中，从而支持在整篇文档和单个数据块层面进行检索。这种双重索引策略能够提升精确率和可解释性，让用户了解 ARC 元数据的哪些层级与其查询最相关。以往的语义搜索系统也采用过类似方法来提升检索粒度，并增强用户对模型输出的信任 [22]。为支持这一方法，每个嵌入都会连同表 I 中所示的信息一并存储，用于在检索界面中提供结合上下文的重点标注，并根据匹配内容的类型辅助筛选和排序。文档和结构的索引可扩展至纳入任何新生成的 ARC 元数据。

5) 混合评分： 匹配用户查询中的关键词对于科学数据探索可能极具价值，尤其是在处理结构化或形式化元数据时。对于简短、部分结构化，或完全由技术术语组成而非完整自然语言句子的输入查询，这一点尤为重要。为此，我们采用一种平衡语义相似度和词汇相似度的混合评分模型 [23]。最终相关性得分通过结合基于 FAISS 的归一化余弦相似度与 BM25 分数得出。两个分数都会先缩放到 0 到 1 之间，然后通过加权策略进行合并。如公式 1 所示，该评分模型既用于完整文档检索，也用于部分数据块检索。

$$S_{\text{final}} = \alpha \cdot S_{\text{D}} + (1 - \alpha) \cdot S_{\text{cmax}} \tag{1}$$

其中， S_{D} 表示查询向量与 ArcBERT 文档嵌入之间的加权相似度得分，而S_{cmax} 表示所有分块中的最大加权得分。权重参数 α可进行调整，以优先考虑文档得分或数据块得分。此外，当用户指定某些元数据结构筛选器（例如调查标题、测定或研究）时，最终得分S_{final} 还可以得到提升。

6) 检索后摘要：系统提供灵活的界面，允许用户以交互方式定义自然语言查询。查询处理层生成查询嵌入，执行语义匹配，并使用混合评分模型对排名前 5 的结果进行排序。系统以自然语言总结检索到的文档，帮助用户通过连贯的概述理解内容，从而提升易用性。在

检索流程的最后阶段，我们集成谷歌的 Gemini API，为每个排名靠前的结果（top-k）生成简洁、结构化的摘要。这些摘要仅用于帮助解读结果，并不会影响其排序。

7）元数据更新：我们使用注册表应用程序中的最新ARC元数据语料库快照优化 ArcBERT 模型时，新创建的 ARC 文档可以直接处理为文档嵌入和文本块嵌入，以便建立索引。系统将继续纳入新的 ARC，以进行语义匹配。不过，必须简化更新流程，确保模型定期使用最新的元数据语料库进行微调。

V. 实验

我们评估 ArcBERT 在元数据探索方面的效果，并将其与传统文本搜索引擎 ElasticSearch 进行比较。我们的主要目标是评估检索结果中排名前 5 项的相关性及其排序。

A. 设置

所有实验均在一台运行 Ubuntu 22.04.5 LTS 的专用机器上进行。服务器配备 AMD EPYC 7662 64核处理器和 1 TB 内存。模型训练任务由英伟达 GPU 加速，具体使用 A100 PCIe，配备 80 GB显存。开发环境基于 Python 3.10，并采用 PyTorch 2.0 作为主要深度学习框架。

对于句子级语义检索，我们实现了 sentence-transformers 库（v2.2.2），该库基于 Hugging Face Transformers 构建。我们使用 FAISS（v1.7.3）在高维空间中进行近似最近邻搜索，而 ElasticSearch（v8.9.1）则作为基于关键词的检索引擎基线。

ArcBERT 基于 all-mpnet-base-v2 架构，并使用结构化 ARC元数据进行训练。我们通过 MultipleNegatives Ranking Loss（MNRL）函数对其进行优化，以提升嵌入空间中的语义连贯性。此外，我们通过官方 Google Generative AI客户端调用 Gemini摘要模块，并使用 Gemini-1.5-flash 模型。

B. 基准查询集

我们整理了一组包含 122 个测试查询的数据集，用于评估 ArcBERT 相较于 ElasticSearch 的有效性。这些查询按照逻辑划分为九个不同类别，分别代表不同的用户查询需求。查询内容从简单的关键词搜索（缩写为 KW），到更复杂的查询不等，后者会纳入分层元数据，例如参数（KWPAR）、研究（KWSUD）、测定（KWA），或以自然语言进行的查询（SWK 和 SEM）。这组查询旨在模拟现实场景：用户可能从不同角度探索数据，例如查找特定的实验技术，或参考已知的贡献者或出版物。查询示例如下：

- 语义型： 哪项 *ARC* 研究探讨了番茄和马铃薯等作物中的抗病基因如何持续活跃?
- 使用关键词：哪项 *ARC* 研究探讨了番茄和马铃薯中抗性基因的全局表达模式？

C. 查询执行

每条查询都会分别在两个检索系统上执行，并根据匹配分数选取前五条结果，再按查询类别计算平均值。值得注意的是，纳入人员、参数、测定和研究等结构化字段后，便可进行分层筛选，让用户进一步细化搜索结果。这些筛选条件模拟了用户在探索复杂元数据时缩小查询范围的方式。

我们使用多种检索指标，对各查询类别对应的结果集取平均后进行比较，包括平均排名、平均倒数排名（MRR），以及查询与结果之间的相似度分数。这些指标用于评估排序精确度和结果的语义对齐情况。MRR 突显了尽早检索到相关文档的重要性，而平均排名则从更宏观的角度展示结果在高位排名中的分布情况。

采用一致的评分值归一化策略，对于确保两个检索系统之间的比较具有意义至关重要。ElasticSearch 使用 BM25 排名算法，该算法生成无界输出分数，而这些分数会根据查询词在索引语料库中的频率和分布而显著变化。若不进行归一化，这些变化可能导致定量评估产生误导，尤其是在使用平均倒数排名（MRR）等共享指标比较性能，或评估排序结果的质量时。这种方法与大规模信息检索（IR）任务（如 TREC 和 CLEF）中采用的评估方法一致：在这些任务中，不同系统使用共享的基于排名的指标进行基准测试，而非依赖各自原生的评分输出。为此，我们对 ElasticSearch 分数采用全局最小-最大归一化策略，如下所述：

$$\text{Normalized Score}_{\text{ES}} = \frac{S - S_{\min}}{S_{\max} - S_{\min}} \tag{2}$$

D. 结果

1) 平均检索分数：图 4 显示，在多种查询类别中，ArcBERT 的前 5 条结果平均相似度分数始终较高。这些结果表明，ArcBERT 等语言模型在理解语义丰富的自然语言查询方面具有显著优势，因此能够提供与用户意图更为契合的结果集。对于包含特定关键词的查询，ElasticSearch 能够高效地检索出匹配度最高的首条结果；但在考量前 5 条结果的平均分数时，其效果则较不理想。这一发现符合预期，因为当用户查询与索引元数据字段高度吻合时，基于词元的检索系统通常表现更佳。有趣的是，ArcBERT 的

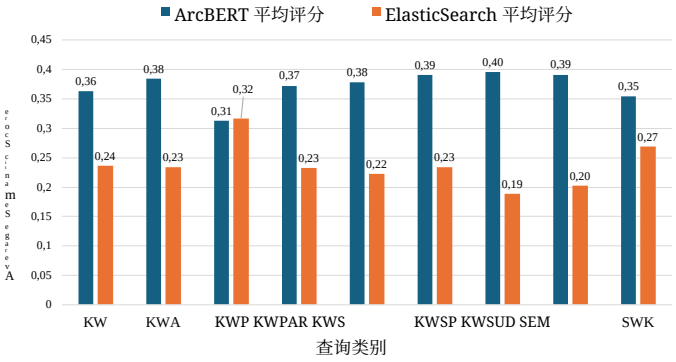

图4. ArcBERT 和 ElasticSearch 在各查询类别中的平均前5名相似度分数。

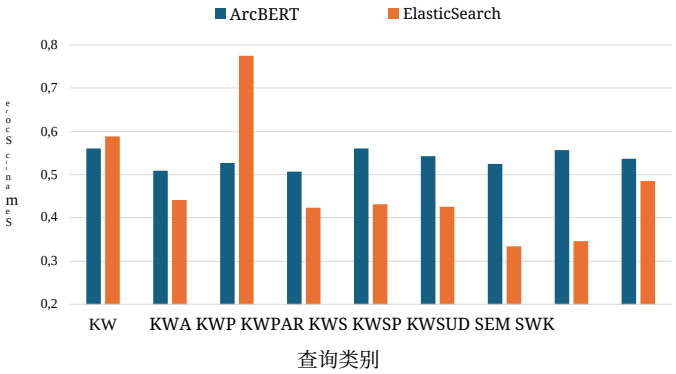

图5. 各查询类别中每个查询的排名第1结果的平均相似度分数。

混合方法结合了余弦相似度和 BM25 词法评分，使其在不同查询类别的前 5 条结果中表现得更加稳定。

2）前几名结果的平均分数：图5展示了 ArcBERT 和 ElasticSearch 在不同查询类别中检索到的排名最高结果的平均分数。对于查询与 ARC 元数据没有显著词汇重叠的类别，ArcBERT 凭借其语义编码能力，始终取得显著更高的首位排名分数。相比之下，ElasticSearch 在这些类别中的首位排名分数表现不佳。然而，在 KW、KWP 和 KWA 等关键词密集型类别中，ElasticSearch 的表现相当，甚至更胜一筹，因为它能够利用精确的词元匹配。

3）平均倒数排名：平均倒数排名（MRR）是一项衡量系统在检索文档列表中将最符合预期的结果排在多靠前位置的指标。图6比较了不同查询类别的 MRR。不出所料，ElasticSearch 在所有类别中的表现均优于 ArcBERT，展现出其将最佳匹配文档排在前列的能力。相比之下，ArcBERT 评分模型更注重稳定的语义匹配。

VI. 结论与未来工作

在本文中，我们介绍了 ArcBERT，这是一款基于 Sentence-BERT 的语义搜索引擎，旨在帮助用户

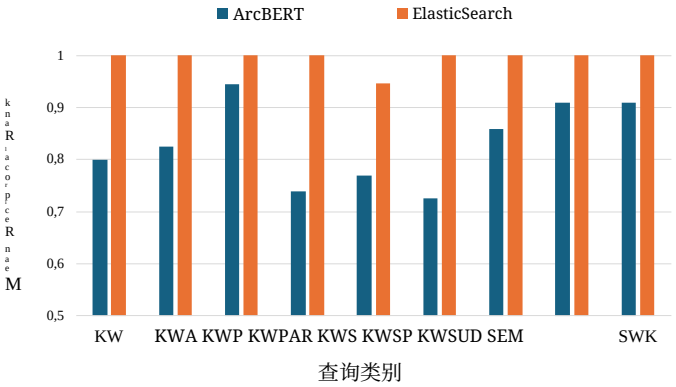

图 6. 各查询类别的 MRR 对比。

通过自然语言查询探索组学元数据。ArcBERT 经过预训练和微调，能够理解植物科学术语以及整合组学元数据的各个层次。我们使用一个模拟真实用户查询模式的查询数据集，将 ArcBERT 与传统文本搜索引擎进行比较，以评估其有效性。正如预期，我们的实验结果表明，ArcBERT 能够理解自然语言风格的查询和元数据的结构；在检索与用户查询语义上更接近且一致的结果方面，它优于文本搜索引擎。不过，我们也观察到，文本搜索引擎仍能通过关键词匹配有效检索出匹配度最高的结果。

ArcBERT 在易用性方面具有显著优势，因为它能够理解自然语言查询。允许用户使用通俗英语编写查询，在生物科学领域尤为宝贵，因为以关键词为中心的数据检索发挥着重要作用。未来，我们计划微调其混合评分函数，以便更好地平衡语义查询匹配与关键词查询匹配。此外，在将 ArcBERT 集成到现有的 ARC 元数据注册库应用程序之前，我们还计划简化使用新元数据重新训练模型的流程。

致谢

我们感谢 DataPLANT（NFDI 7/1 – 42077441）提供的支持。DataPLANT 是德国国家研究数据基础设施的一部分，该基础设施由德国研究联合会（DFG）资助。

参考文献

[1] M. Krantz、D. Zimmer、S. O. Adler、A. Kitashova、E. Klipp、T. M¨uhlhaus 和 T. N¨agele，《植物生物学中的数据管理与建模》，《植物科学前沿》，第 2021 [ ] 卷。在线获取。可访问：https://www.frontiersin.org/journals/plant- science/articles/ 10 3389 2021 717958 。/fpls。

[2] C. Pommier、F. Coppens、H. Cwiek-Kupczy´nska、D. Faria、S. Beier、C. Miguel、C. Michotey、F. D’Anna、S. Owen 和 K. Gruden，植物科学数据集成：从构建社区标准到定义一致的数据生命周期。沙姆：施普林格国际出版，2023 年，第 149–160 页。 [在线]。可访问：https://doi.org/10.1007/978-3-031-13276-6 8

[3]K. Haug、R. M. Salek 和 C. Steinbeck，《代谢组学中的全球开放数据管理》，《化学生物学当前观点》，第 36 卷，第 58–63 页，2017 年，组学。 [在线]。可访问：https://www.sciencedirect.com/science/article/pii/S1367593116302083

[4] W. Wruck、M. Peuker 和 C. R. Regenbrecht，《跨国大型系统生物学项目的数据管理策略》，《生物信息学简报》，第 15 卷，第 1 期，第 65–78 页，2012 年 10 月。 [ ] 10 1093 064 在线。可访问：https://doi.org/。/bib/bbs K. Haug、R. M. Salek、P. Conesa、J. Hastings、P. de Matos、M. Ri-j nbeek、T. Mahendraker、M. Williams、S. Neumann、P. Rocca-Serra、E. Maguire、A. Gonz´alez-Beltr´an、S.-A. Sansone、J. L. Griffin 和 C. Steinbeck，《MetaboLights——用于代谢组学研究及相关元数据的开放获取通用型存储库》，《核酸研究》，第 41 卷，数据库专刊，第 D781–6 页，2012 年 10 月。[6] K. Haug、K. Cochrane、V. C. Nainala、M. Williams、J. Chang、K. V. Jayaseelan 和 C. O'Donovan，《MetaboLights：一个响应科学界需求而不断发展的资源》，《核酸研究》，第 48 卷，第 D1 期，第 D440–D444 页，2019 年 11 月。 [在线]。可访问：https://doi.org/ 10 1093 1019 。/nar/gkz H. Hermjakob 和 R. Apweiler，《蛋白质组学鉴定数据库（PRIDE）与 ProteomExchange联盟：让蛋白质组学数据可访问》，《蛋白质组学专家评论》，第 3 卷，第 1 期，第 1–3 页，2006 年，pMID：16445344。 [在线]。可访问：https://doi.org/ 10 1586 14789450 3 1 1 。/。。。H. Y. I. Lam、X. E. Ong 和 M. Mutwil，《植物生物学中的大语言模型》，CoRR，第 abs/2401.02789 卷，2024 年。 [在线]。可访问：https://doi.org/ 10 48550 2401 02789 。/arXiv。。。Q. Zhang、K. Ding、T. Lyv、X. Wang、Q. Yin、Y. Zhang、J. Yu、Y. Wang、X. Li、Z. Xiang、X. Zhuang、Z. Wang、M. Qin、M. Zhang、J. Zhang、J. Cui、R. Xu、H. Chen、X. Fan、H. Xing 和 H. Chen，《科学大语言模型：生物与化学领域综述》，CoRR，第 abs/2401.14656 卷，2024 年。 [在线]。可访问：https://doi.org/ 10 48550 2401 14656 。/arXiv。。。J. Lee、W. Yoon、S. Kim、D. Kim、S. Kim、C. H. So 和 J. Kang，《BioBERT：用于生物医学文本挖掘的预训练生物医学语言表征模型》，CoRR，第 abs/1901.08746 卷，2019 年。 [在线]。可访问：http://arxiv.org/abs/ 1901 08746 。I. Beltagy、K. Lo 和 A. Cohan，《SciBERT：用于科学文本的预训练语言模型》，载于《2019 年自然语言处理实证方法会议与第九届自然语言处理国际联合会议论文集》，EMNLP-IJCNLP 2019，中国香港，2019 年 11 月 3—7 日，K. Inui、J. Jiang、V. Ng 和 X. Wan 编。计算语言学协会，2019 年，第 3613 3618 [ ] 10 18653 1 19 1371 – 页。在线。可访问：https://doi.org/。/v /D - J. Devlin、M. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，载于《2019 年计算语言学协会北美分会会议：人类语言技术论文集》，NAACL-HLT 2019，2019 年 7 月 2—1 日，美国明尼阿波利斯，第一卷（长篇与短篇论文），J. Burstein、C. Doran 和 T. Solorio 编。计算语言学协会，2019 年，第 4171 4186 [ ] 10 18653 1 19 1423 – 页。在线。可访问：https://doi.org/。/v /n - N. Reimers 和 I. Gurevych，《Sentence-BERT：使用孪生 BERT 网络的句子嵌入》，载于《2019 年自然语言处理实证方法会议与第九届自然语言处理国际联合会议论文集》，EMNLP-IJCNLP 2019，中国香港，2019 年 11 月 3—7 日，K. Inui、J. Jiang、V. Ng 和 X. Wan 编。计算语言学协会，2019 年，第 3980–3990 页。 [在线]。可访问：https://doi.org/ 10 18653 1 19 1410 。/v /D - 在线，《PubMed》，https://pubmed.ncbi.nlm.nih.gov/，网页；访问日期：2025-04-14。[15]H. L. Weil、K. Schneider、M. Tsch¨ope、J. Bauer、O. Maus、K. Frey、D. Brilhaus、C. Martins Rodrigues、G. Doniparthi、F. Wetzels、J. Lukasczyk、A. Kranz、B. Gr¨uning、D. Zimmer、S. Deßloch、D. von Suchodoletz、B. Usadel、C. Garth 和 T. M¨uhlhaus，《PLANTdataHUB：用于植物研究中持续开展 FAIR 数据共享的协作平台》，《植物学杂志》，第 116 卷，第 4 期，第 974–988 页，2023 年。 [在线]。可访问：https://onlinelibrary.wiley.com/doi/abs/ 10 1111 16474 。/tpj。C. Garth、J. Lukasczyk、T. Muhlhaus、B. Venn、J. Kr¨uger、K. Glogowski、C. Martins Rodrigues 和 D. von Su-chodoletz，《不可变但不断演进：研究数据—时间连续体中用于永久共享的 ARC》。heiBOOKS，2022年4月，第366–373页。 [在线]。网址：https://books.ub.uni-heidelberg.de/heibooks/catalog/book/ 979 13751 /chapter/

[17] DataPLANTcommunity，《注释研究上下文规范，v1.1-rfc》，2023年8月。[在线]。网址：https://doi.org/ 10 5281 8302662 。/zenodo。S.-A. Sansone、P. Rocca-Serra、A. Gonzalez-Beltran、D. Johnson 和 I. Community，《*ISA* 模型与序列化规范 *1.0 - ISA JSON* 格式》，https://isa-specs.readthedocs.io/en/latest/isajson.html，2016，[2024 02 07] 网页；访问日期：--。[19] Soiland-Reyes, Stian 、 Sefton, Peter 、 Crosas, Merc`e 、 Castro, Leyla Jael、Coppens, Frederik、Fern´andez, Jos´e M、Garijo, Daniel、Gr¨uning, Bj¨orn、La Rosa, Marco、Leo, Simone 等，《使用 RO-Crate 封装研究成果》，数据科学，第5卷，第2期，第97–138页，2022年。[20] T. B. Brown、B. Mann、N. Ryder、M. Subbiah、J. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell、S. Agarwal、A. Herbert-Voss、G. Krueger、T. Henighan、R. Child、A. Ramesh、D. M. Ziegler、J. Wu、C. Winter、C. Hesse、M. Chen、E. Sigler、M. Litwin、S. Gray、B. Chess、J. Clark、C. Berner、S. McCandlish、A. Radford、I. Sutskever 和 D. Amodei，《语言模型是少样本学习器》，载于神经信息处理系统进展 *33*：*2020*年神经信息处理系统年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan 和 H. Lin 编，2020年。[21] P. P. Ghadekar、S. Mohite、O. More、P. Patil、Sayantika 和 S. Mangrule，《使用 FAISS 的句子语义相似度检测器》，载于*2023*年第*7*届国际计算、通信、控制与自动化会议 *(ICCUBEA)*，2023年，第1–6页。[22] V. Karpukhin、B. Oguz、S. Min、P. S. H. Lewis、L. Wu、S. Edunov、D. Chen 和 W. Yih，《用于开放域问答的稠密段落检索》，载于*2020*年自然语言处理实证方法会议论文集，*EMNLP 2020*，线上，*2020*年*11*月*16*日至*20*日，B. Webber、T. Cohn、Y. He 和 Y. Liu 编。计算语言学协会，2020年，第6769–6781页。 [ ] 10 18653 1 2020 550 在线。网址：https://doi.org/。/v / .emnlp-main。A. Askari、A. Abolghasemi、G. Pasi、W. Kraaij 和 S. Verberne，《将 BM25 分数作为文本输入可提升基于 BERT 的重排序器效果》，载于信息检索进展——第*45*届欧洲信息检索会议，*ECIR 2023*，爱尔兰都柏林，*2023*年*4*月*2*日至*6*日，论文集，第一部分，丛书：计算机科学讲义，J. Kamps、L. Goeuriot、F. Crestani、M. Maistro、H. Joho、B. Davis、C. Gurrin、U. Kruschwitz 和 A. Caputo 编，第13980卷。Springer，2023年，第66–83页。 [在线]。网址：https://doi.org/10.1007/978-3-031-28244- 7 5