arXiv:2601.01118v1 [cs.IR] 3 Jan 2026

ScienceDB AI：面向大规模科学数据共享服务的大语言模型驱动的智能体推荐系统

Qingqing Long、Haotian Chen、Chenyang Zhao、Xiaolei Du、Xuezhi Wang、Pengyao Wang、Chengzan Li、Yuanchun Zhou*、Hengshu Zhu* 中国科学院计算机网络信息中心，中国北京

摘要

科学智能（AI4S）的快速发展凸显了科学数据集的重要性，推动了众多国家科学数据中心和共享平台的建立。尽管取得了这些进展，高效推动数据集共享及其在科学研究中的利用仍面临挑战。科学数据集包含复杂的领域专属知识和背景信息，因此传统的基于协同过滤的推荐系统难以胜任。近年来，大语言模型（LLM）的进展为构建能够深入理解语义并提供个性化推荐的对话代理带来了前所未有的机遇。为此，我们推出了**ScienceDBAI**，这是一个基于科学数据银行（ScienceDB）开发的新型大语言模型驱动的智能体推荐系统。科学数据银行是全球最大规模的科学数据共享平台之一。ScienceDB AI 利用自然语言对话和深度推理，准确推荐符合研究人员科学意图及其不断变化需求的数据集。该系统带来了多项创新：科学意图感知器，可从复杂查询中提取结构化实验要素；结构化记忆压缩器，可有效管理多轮对话；以及可信检索增强生成（Trustworthy RAG）框架。可信 RAG 采用两阶段检索机制，并通过可引用科学任务记录（CSTR）标识符提供可引用的数据集引用，从而提升推荐可信度和可复现性。通过使用超过 1000 万个真实世界数据集开展大量离线和在线实验，ScienceDB AI 展现出显著效果：与先进基线方法相比，离线指标提升超过 30%；与基于关键词的搜索引擎相比，点击率提高超过 200%。据我们所知，ScienceDB AI 是首个专为大规模科学数据集共享服务打造的、由大语言模型驱动的对话式推荐系统。该平台可公开访问

访问：https://ai.scidb.cn/en。

关键词

数据集推荐、科学数据、对话式推荐、代理推荐系统、数据共享服务、大语言模型

*通讯作者。

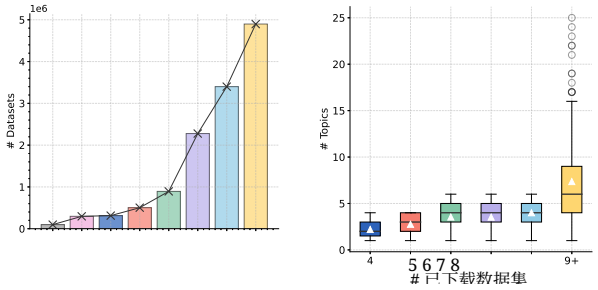
允许免费为个人或课堂使用而制作本作品的全部或部分数字或纸质副本，但前提是副本不得用于或分发以获取利润或商业利益，并且副本首页须附有本声明及完整引用。对于本作品中由作者以外的其他人拥有版权的组成部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、转载、发布到服务器或分发至列表，则必须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。会议简称 *'XX, xx, xx* © 2026年 版权归所有者/作者所有。出版权已授予 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM https://doi.org/XX

ACM 参考格式：

Qingqing Long、Haotian Chen、Chenyang Zhao、Xiaolei Du,, Xuezhi Wang、Pengyao Wang、Chengzan Li、Yuanchun Zhou[1], Hengshu Zhu。2026年。Sci- enceDB AI：面向大规模科学数据共享服务的大语言模型驱动的智能体推荐系统。载于论文集。请根据权利确认邮件填写正确的会议标题（会议简称 *'XX*）。ACM，美国纽约州纽约市，12页。https://doi.org/XX

1 引言

科学人工智能（AI4S）的快速发展 [6,21, 31, 61] 凸显了高质量科学数据对于加速生物学、物理学、化学和地球科学等领域发现的关键作用 [21, 37, 44, 48, 59], 。为此，世界各地的政府和研究机构建立了国家科学数据中心 [12] 和数据集共享平台，例如 NCBI [12]、 OpenAIRE [42] 和 ScienceDB [60]。这些举措推动开放获取并促进科学数据的协作利用，从而提高其可再用性。因此，近年来新发布的科学数据集数量显著增加 [43, 46, 47], ，如图 1(a) 所示。



（a）Zenodo [43], 、ScienceDB [60] 和 FigShare [46] 上每年发布的科学数据集数量。

（b）ScienceDB 中不同研究者行为长度下的数据集学科分布。

图 1：数据集和用户行为的统计结果。

随着科学数据集的迅速增长，帮助研究人员高效地发现相关数据集变得愈发重要。因此，有效的数据集推荐系统对于促进数据驱动的科学发现至关重要 [1, 3, 47]。传统的数据集推荐系统通常分为两类。第一类是基于行为的推荐系统，通过协同过滤（CF） [30, 56, 57] [8, 19, 和图表 27, 29] 示学习（GRL）等方法利用用户交互历史。第二类是基于内容的推荐系统，依赖查询本身，包括基于关键词的检索 [47, 60] 和基于语义嵌入的匹配 [1, 3, 28]。现有的数据集共享平台，如 Google Dataset Search [5], 、DataCite

Commons [38], 、OpenAIRE [42] 和 Dryad [16], 等，仍然都高度依赖基于关键词的搜索引擎。其详细信息见表 1。尽管这些工作已取得一定成功，但大规模科学数据集推荐仍面临**尚未得到充分解决的独特挑战**：

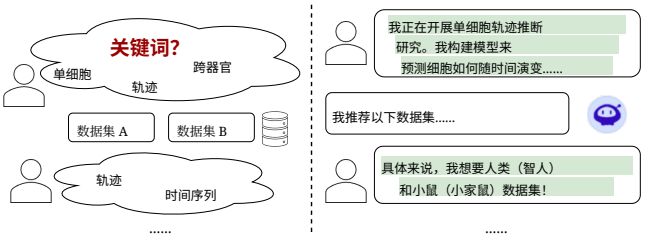


图 2：我们研究动机的示意图。左图展示了现有数据集共享平台面临的挑战。右图说明了我们的 ScienceDB AI 能够深入理解研究人员的实验数据集需求。

(1) 科学数据集需求通常**与具体任务相关，且与历史行为无关**。如图 1（b）所示，x 轴表示研究人员此前在 ScienceDB [60], 下载的数据集数量，y 轴表示涉及的不同主题数量。下载次数为 ≥9 （约占总人数的 10%）的研究人员被归为一组。图中显示，研究人员的下载历史在主题方面的一致性较弱，这意味着他们的数据集需求更多由不断变化的研究任务驱动，而非持久的偏好。因此，基于用户行为的推荐系统不适用于我们的场景。

(2) 现有基于上下文的推荐系统**难以理解实验层面的数据集需求**。科学探索通常涉及高度具体、不断变化且细致入微的数据集需求，并通过丰富的自然语言描述来表达。传统的关键词搜索或基于嵌入的匹配难以理解这些复杂需求。例如，如图 2 所示，研究人员可能会查询：“我正在开展一项关于单细胞命运轨迹推断的研究，重点关注人类（智人，*Homo sapiens*）和小鼠（小家鼠，*Mus musculus*）的跨器官分化。我希望构建模型，预测单个细胞如何随时间演变以及如何响应遗传扰动”。这类详尽且具有领域特异性的意图需要深度的上下文理解，而现有的基于上下文的推荐系统尚无法有效处理。

幸运的是，大语言模型和智能体在对话式推荐领域的最新进展，为解决我们的问题提供了一个有前景的方向 [25]。然而，这些模型本质上**容易出现幻觉和遗忘问题** [14, 21]。它们可能生成虚构的、不存在的或无法访问的数据集。这在科学场景中构成了关键挑战，因为可信、可访问且可引用是基本要求 [11, 39]。

针对这一问题，我们提出了 ScienceDB AI，这是一种面向大规模科学数据共享服务设计的智能体推荐系统。我们的系统基于一个包含超过 1,000 万个可用数据集的资源库运行，并引入了多个关键组件，以支持可信、易用且可引用的数据集推荐。首先，我们开发了一个实验意图感知器，将研究人员的数据、主题、约束条件和评估标准提取为结构化意图模板。其次，我们引入了结构化记忆压缩器。它会跟踪多轮对话中的用户意图、对话上下文和工具调用，并总结相关历史信息。这有助于缓解大语言模型上下文窗口有限所导致的遗忘问题。第三，为了解决幻觉问题，我们提出了可信检索增强生成（Trustworthy RAG）框架。该框架采用两阶段检索器，在大规模场景下平衡检索效果与效率。为确保数据集的可追溯性和引用，我们为每个数据集关联一个可引用科学任务记录（CSTR），并在系统响应中提供指向 CSTR 的直接链接。我们基于 ScienceDB 平台中超过 1,000 万个真实世界科学数据集开展了广泛的离线和在线评估。ScienceDB AI**的离线指标较现有基于智能体的推荐系统提升了 30% 以上**。在**在线A/B测试中，其点击率（CTR）较传统基于关键词的搜索系统提升了 200% 以上**。我们的贡献总结如下：

- 据我们所知，ScienceDB AI 是首个面向大规模科学数据共享服务的、由大语言模型驱动的智能体推荐系统。
- 我们设计了一个智能体框架，由实验意图感知器、结构化记忆压缩器，以及一个为每个数据集附加 CSTR 以确保可信度的检索增强型推荐系统组成。
- 通过在 1000 多万个真实世界数据集上进行大量实验，ScienceDB AI 的离线指标显著提升（30%+），在线 A/B 测试也取得了显著增长（200%+）。

2 相关工作

本节首先回顾现有的科学数据集共享平台，重点介绍其优势与局限性。随后，我们考察旨在促进数据集可发现性的 数据集推荐系统研究。最后，我们讨论基于代理的对话式推荐系统的最新进展。

2.1 科学数据集共享平台

AI4S 的最新进展凸显了高质量科学数据的重要性 [44, 48]。全球各地的政府和研究机构已建立国家科学数据中心和数据集共享平台。在此，我们从五个维度对 14 个现有数据集共享平台进行比较：(1) 支持的学科数量，(2) 是否针对研究应用场景设计，(3) 是否提供源数据，(4) 可用数据集的数量，以及 (5) 是否提供对话式推荐系统（CRS）。学科数量根据 OpenAlex [4] 的主要学科分类体系进行估算。详细比较见表 1。Snowflake Market- place [2] 和 DataBricks [32] 是两款商业产品，因此其数据集信息未知。如表所示，半数平台支持约 10 个学科，其余平台支持的学科少于五个。相比之下，我们的平台覆盖 18 个一级学科，主题覆盖面更广，且提供更多样化的领域专属数据集。在所有平台中，Google Dataset Search [5], ScienceDB [60] 和 DataCite Commons [38] 托管的数据集数量最多

表 1：数据集共享服务比较。

产品/平台	# 学科	用于研究	共享 SourceData	# 数据集	CRS
DataCite Commons [38]	>10	✓	✓	42,896,080	✗
Google Dataset Search [5]	>10	✓	✗ (仅元数据)	2500 万	✗
Zenodo [43]	<5	✓	✓	400 万	✗
OpenAIRE [42]	>10	✓	✓	8,382,956	✗
飞桨 [7]	>10	✓	✓	~10,000	✗
Dataverse [33]	>10	✓	✓	139,231	✗
CKAN [52]	>10	✓	✓	24,233	✗
Dryad [16]	<5	✓	✓	~900,000	✗
Snowflake Marketplace [2]	未知	✗ (商业)	✓	未知	✗
DataBricks [32]	未知	✗ （商业用途）	✓	未知	✗
HuggingFace [18]	<3	✓	✓	461,199	✗
RADx 数据中心 [35]	<2	✓	✓	~5,000	✗
NCBI [12]	<3	✓	✓	~1,000	✗
FigShare [46]	<3	✓	✓	~380,000	✗
ScienceDB AI（我们的系统）	**全部（>18）**	✓	✓	**1000 万**	✓

然而，Google Dataset Search 仅对元数据进行索引，不提供源数据，因此限制了其在实验研究中的适用性。

总而言之，现有数据平台缺乏对数据集共享和推荐的有效支持。相比之下，ScienceDB AI 是唯一能够提供智能推荐的数据中心，使研究人员能够用自然语言表达复杂的数据需求，并高效地发现相关数据集，最终加速科学发现。

2.2 数据集推荐系统

近年来，只有三项具有代表性的研究专门针对数据集推荐任务。DataFinder [47] 提出了一种基于文本相似度的数据集推荐模型。该模型使用 BERT 作为数据集描述和用户输入查询的嵌入模型。Altaf 等人 [1] 提出了一种用于基于查询的数据集推荐任务的变分图自编码器。他们构建了一组研究论文，用以反映用户的研究兴趣。推荐的数据集基于数据集描述的表示与为该用户构建的研究论文图之间的相似度。DataLinking [3] 使用概念频率和词频-逆文档频率提取用户查询与数据集描述的相似度特征。然而，这些研究主要基于关键词，既无法理解研究人员的复杂需求，也不支持交互式自然语言查询。

2.3 基于智能体的对话式推荐系统

许多研究表明，[50, 54] 大语言模型和基于智能体的对话式推荐系统在理解用户复杂意图方面的表现优于传统模型。它们能够利用专用工具，从而缓解模型规模和预训练数据量受限所导致的知识不足。代表性工作包括 AgentCF [56], InteRecAgent [17] 和 CoSearchAgent [13],等。具体而言，CoSearchAgent、Fang 等 [10], 和 MACRec [50] 属于多智能体协同搜索系统。然而，多智能体系统存在通信延迟，导致系统响应时间较长，因此难以适用于大规模在线推荐场景。因此，本研究关注单智能体推荐工作。AgentCF 设计了基于智能体的协同过滤，以模拟用户-项目交互。InteRecAgent、ChatCRS [23], 和 RecMind [49] 设计了基于智能体的对话框架，其中包含规划、记忆、网络搜索、反思和推荐工具等机制。其他基于智能体的工作 [15, 51, 53, 58] 主要关注对话中的个性化推荐。

然而，以上所有模型都天然容易产生幻觉 [21], ，常会为不存在或无法访问的数据集生成推荐。这在科学场景中构成了一项重大挑战，因为可信、可访问且可引用是基本要求 [11, 39]。此外，这些模型主要是基于行为的模型，因此不适合理解实验层面的查询。

3 ScienceDBAI 的技术细节

本节将详细介绍 ScienceDB AI 的技术细节。首先，我们概述技术框架并定义问题。随后，我们介绍框架的组成部分，即实验意图感知器、结构化记忆压缩器，以及一种检索增强型推荐系统；该系统为每个数据集附加唯一标识符，以确保可信度。

3.1 框架概述

框架概述。Sci- enceDB AI 的整体技术框架如图 3 所示，由实验意图感知器、结构化记忆压缩器和检索增强型推荐系统组成；该推荐系统为每个数据集附加唯一标识符，以确保可信度。可访问我们的 ScienceDB AI 在线系统：https://www.ai.scidb.cn/en。我们的在线网页示例如图 4 所示。

问题定义。令 $Q = \{q_1, q_2, ..., q_T\}$ 表示多轮研究人员查询，其中 q_t 表示第 t 轮输入查询，

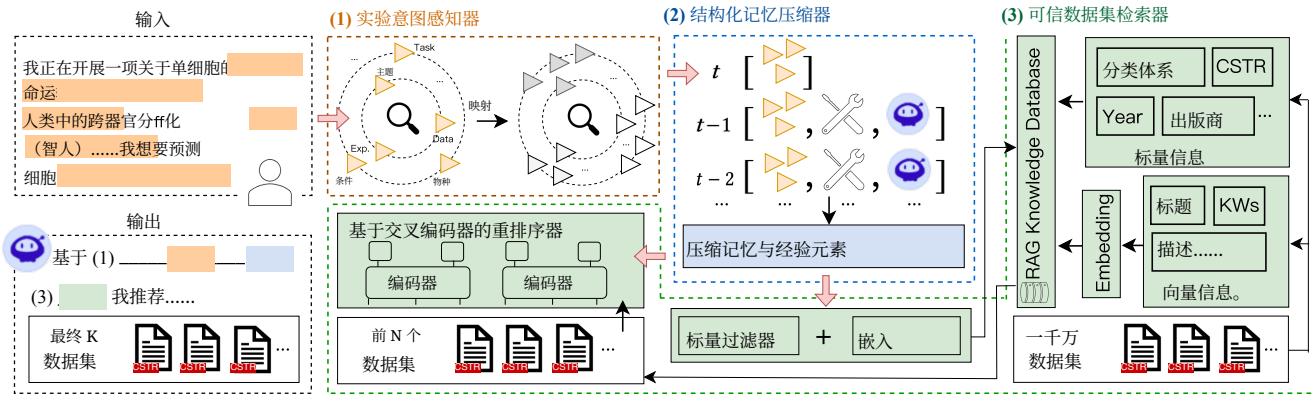


图3：我们设计的 ScienceDB AI 系统技术框架。系统由实验意图感知器、结构化记忆压缩器和检索增强型推荐系统组成，后者为每个数据集附加 CSTR [60] 以确保可信度。

该查询使用技术描述，并包含研究目标、方法描述、实验约束等。令 $D = \{d_1, d_2, ..., d_N\}$ 表示大规模候选数据集，其中本文中的 N 大于 1000 万。每个数据集 d_i 都有相应的元数据信息和文本描述。本文旨在设计一个数据集推荐系统 F，使其能够通过尽可能少的对话，为研究人员推荐最合适的 K（$K \ll N$）个数据集，即尽可能减小 T。

3.2 实验意图感知器

如图2和图9所示，研究人员的实验输入可能极其复杂。为支持面向研究人员的实验级数据集推荐，我们设计了一个实验意图感知器，可将研究人员的长篇自然语言内容提取为**结构化实验要素**。与传统的数据集推荐模型 [1, 3, 47] 和通用推荐系统 [13, 17], 相比，本文旨在构建一款面向对话场景的数据集推荐系统，专为科学研究场景设计。

意图感知器的设计基于科学发现的结构化要素体系和典型流程 [34, 36, 40]。具体而言，数据、主题、实验约束/设置和评估指标是典型的顶层要素。物种和数据模态、来源及注释是数据的典型次级要素。以图3中的输入查询为例，我们的意图感知器识别出研究主题为人类跨器官细胞分化，任务为单细胞命运轨迹推断，且细胞随时间演变。跨器官范围和人体组织背景被解读为实验约束。查询中提取出的科学意图将被改写为 $\tilde{q}_t$。

3.3 结构化记忆压缩器

由于研究人员的需求复杂，其请求可能较长，且与一般推荐任务相比，通常需要更多轮对话。为了有效支持科学场景中的多轮复杂查询，我们设计了一种结构化记忆压缩器，可提炼关键信息并在此过程中保留依赖上下文的依赖关系。该模块旨在应对大语言模型固有的遗忘问题 [14, 21] 。

我们会在平台中跟踪所有实时对话状态和历史记录。令 $\Theta_{1:T}$ 表示截至第 t轮的对话历史，则

$$\Theta_{1:T} = \{(\tilde{q}_1, \tau_1, r_1), \ldots, (\tilde{q}_T, \tau_T, r_T)\}, \tag{1}$$

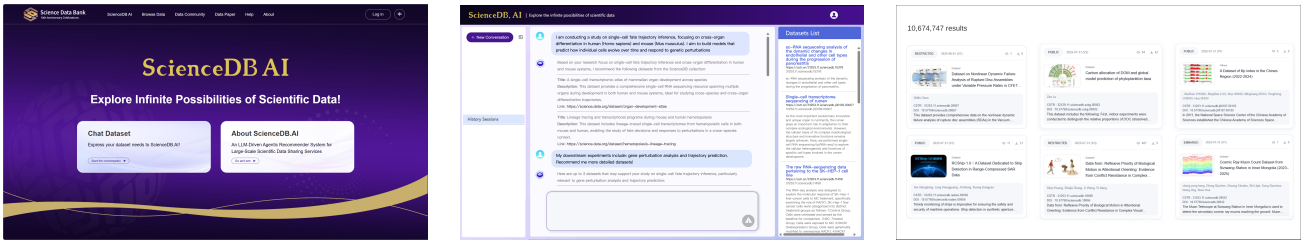
其中，τ_t 表示工具调用和执行日志。工具日志有助于避免在后续对话轮次中执行冗余操作。r_t 表示 ScienceDB AI 在第 t 轮的回复。记忆预算限制为 $L_{\max}$ 词元（例如 32K），因此，完整纳入 $\Theta_{1:T}$ 既有意义，也很有挑战性。如果将极长的对话记录直接输入大语言模型，就会导致大语言模型忘记系统提示词或给定的推荐候选数据集，从而在回复中产生幻觉。本文旨在将 $\Theta_{1:t}$ 压缩为结构化记忆 S_t，保留此前 $t-1$ 轮中的信息，

$$\mathcal{S}_t = \begin{cases} \Theta_{1:1}, & t = 1. \\ \mathcal{M}(\tilde{q}_t, \tau_t, r_t, \mathcal{S}_{t-1}), & t > 1. \end{cases} \tag{2}$$

此外，S_t 还应具备时效性感知冲突消解能力，即优先采用近期更新的信息，而非陈旧或过时的信息。为维护结构化意图模板，我们采用显式压缩，而非隐式压缩 [9] 。M 表示一个基于大语言模型的智能体，用于将历史对话日志总结为结构化信息。随后，压缩后的结构化记忆 S_t 将作为大语言模型生成最终回复的上下文。当因语义歧义而无法解决冲突时，我们会主动提出澄清问题，例如：“您是否希望覆盖之前的数据集约束.....？”

3.4 可信数据集检索器

为获得更准确的检索候选结果，我们在大规模数据集共享服务中采用两阶段检索器，以权衡效果与效率。每个数据集 d_i 都关联着稠密嵌入和结构化元数据，例如发布时间和所属机构（如附录中的图 8 所示）。在第一阶段，我们通过



(a) 入口 (b) 聊天示例 (c) 千万数据集示例

图 4：我们的 ScienceDB AI 在线平台，可通过 https://ai.scidb.cn/en 访问。

结合预过滤的向量相似度检索排名前 d_i 的候选数据集。如果输入查询明确包含某些约束，或大语言模型提取出诸如出版日期、分类体系或所属机构等约束，我们会应用标量过滤，以缩小候选空间。随后，我们计算查询嵌入 $\boldsymbol{e}(\tilde{q}_t)$ 与数据集描述 $\boldsymbol{e}(d_i)$ 之间的余弦相似度，并选出相似度最高的前 N 个数据集。在第二阶段，为了深入理解研究人员意图，我们采用 ColBERT [20] 作为重排序器。重排序器对 $\tilde{\boldsymbol{q}}_t$ 与 N 候选项的词元级嵌入进行细粒度晚期交互，并生成最终排名前 K 的数据集。请注意，最终响应中推荐的数据集数量取决于研究人员的需求。如果输入查询中未特别说明，则将 K 设为 3。

此外，为确保推荐的数据集具备可追溯性和可信度，即能够被唯一识别并可引用，我们为每个数据集附加一个可引用科学任务记录 *(CSTR)* [60]，d_i 并在最终响应中包含相应的 CSTR 链接。CSTR 标识符为科学资源提供唯一且标准化的 ID，类似于 DOI [24]。不过，CSTR 支持更广泛的资源类型。在我们的场景中，它可以唯一识别数据集及其源数据文件，而 DOI 无法做到这一点。具体而言，它有助于消除因数据集及其源文件的名称或存储位置发生变化而产生的歧义。为强制执行这一行为，我们加入如下系统提示词：“对于每个选定的数据集，您必须返回其 *CSTR* 标识。”

我们的 ScienceDB AI 技术框架伪代码如算法1所示。

讨论。 与其他基于大语言模型或智能体的推荐模型相比，我们表明，在大规模数据共享服务中，研究人员意图理解、检索器和记忆模块对满足研究人员的科学需求影响最大，其重要性超过复杂的规划、网络搜索或反思模块。第4.2节提供了支持这一结论的实验证据。

算法1：ScienceDB AI 的算法流程。输入：第 t 轮用户查询 $\boldsymbol{q}_t$；对话历史 $\Theta_{1:t-1}$；包含元数据的数据集索引 D 输出：前 k 个推荐数据集 $\{d_1, \ldots, d_k\}$ 和最终响应 r_t1 初始化：结构化记忆 $S_0 \leftarrow \varnothing$2 步骤 1：实验意图感知器；3 $\tilde{\boldsymbol{q}}_t \leftarrow \text{LLMParse}(\boldsymbol{q}_t, \Theta_{1:t-1})$；// 使用大语言模型解析查询和对话历史，以提取科学意图 4 分解 $\tilde{\boldsymbol{q}}_t = (U, T, D, E, Z)$；// 主题 U、任务 T、数据模态 D、实验设置 E 、评估指标 Z 5 步骤 2：结构化记忆压缩器；6 更新对话日志：$\Theta_{1:t} \leftarrow \Theta_{1:t-1} \cup \{(\tilde{\boldsymbol{q}}_t, \boldsymbol{\tau}_t, \boldsymbol{r}_t)\}$; 7 将 $\Theta_{1:t}$ 压缩为结构化记忆：$S_t \leftarrow \text{SSRC}(\Theta_{1:t})$；// 科学语义保留压缩（SSRC），将历史记录总结为结构化记忆 8 步骤 3：可信数据集检索器；9 嵌入意图：$h_t \leftarrow \text{EmbedIntent}(\tilde{\boldsymbol{q}}_t, S_t)$; 10 嵌入数据集：$H_{\mathcal{D}} = \{h_d \mid d \in D\}$；// 每个 h_d 编码

数据集元数据：描述、关键词、来源等。11 通过近似最近邻（ANN）检索排名前 k 个候选项：

$$\{d_1, \ldots, d_k\} \leftarrow \text{ANN}(\mathbf{h}_t, \mathbf{H}_{\mathcal{D}})$$

12 通过交叉编码器对候选项重新排序：

13 **步骤 4：生成最终响应；14 返回** $\text{score}(\tilde{\boldsymbol{q}}_t, d_i) \leftarrow f_{\text{cross}}(\tilde{\boldsymbol{q}}_t, \text{meta}(d_i))$ $\{d_1, \ldots, d_k\}$ r_t

$$r_t \leftarrow \text{LLMAnswer}(\tilde{\boldsymbol{q}}_t, \mathcal{S}_t, \{(d_i, \text{meta}(d_i))\}_{i=1}^{k}, \text{SystemPrompt})$$

4 实验

本节首先介绍用于评估我们方法的实验设置。随后，我们展示整体性能结果，并分析 ScienceDB AI 的运行效率。接着，我们通过详细的案例研究说明其实际效果。最后，我们报告在线 A/B测试的结果，以全面验证我们的框架。

4.1 实验设置

数据集 我们通过从过去两年 ScienceDB [24, 60] 中的用户-数据集点击日志中随机抽样，构建离线评估数据集。具体而言，我们抽取了约 10,000 名用户及其对应的 15,000 个已下载数据集。对于每位研究人员，**其此前点击的数据集被视为模拟对话中的真实目标**。候选数据集基于

表2：多轮对话数据集推荐的整体性能比较。

模型	召回率			NDCG			MRR			AT		
	@1	@3	@5	@1	@3	@5	@1	@3	@5	@1	@3	@5
DataFinder [47]	0.0115	0.0726	0.1481	0.0115	0.0455	0.0764	0.0115	0.0363	0.0533	3.35	3.09	3.01
DataLinking [3]	0.2605	0.3003	0.3084	0.2605	0.2838	0.2871	0.2605	0.2781	0.2800	3.23	3.06	3.03
DeepSeek-V3 [26]+检索增强生成	0.2277	0.2513	0.2530	0.2277	0.2420	0.2428	0.2277	0.2388	0.2392	3.33	3.24	3.23
通义千问3 [55]+检索增强生成	0.2559	0.2778	0.2824	0.2559	0.2692	0.2712	0.2559	0.2662	0.2673	3.21	3.17	3.15
InteRecAgent [17]	0.2686	0.3083	0.3141	0.2686	0.2926	0.2950	0.2684	0.2871	0.2884	3.20	3.06	3.05
CoSearchAgent [13]	0.1608	0.1988	0.2386	0.1608	0.1822	0.1984	0.1608	0.1766	0.1854	3.41	3.31	3.25
Ours	**0.4064**	**0.4187**	**0.4196**	**0.4065**	**0.4142**	**0.4146**	**0.4065**	**0.4126**	**0.4128**	**3.19**	**2.89**	**2.83**

余弦相似度从 1,000 万个数据集中检索得到。遵循先前的对话式推 [17, 25], 荐工作，我们构建了一个包含多轮交互的离线数据集，以模拟研究人员复杂且专业的需求。为更好地模拟这些复杂性，我们利用大语言模型（Qwen-Plus）基于数据集描述生成实验设计方案。与现有的离线对话式数据集相比，我们的输入查询更加详细、冗长且细致，构成了更具挑战性的对话式推荐任务。对话轮数设置为 3 至 5 轮。详细的离线构建过程和样例见**附录**。

竞争基线。 我们选择以下基线方法作为竞争对象，可将其分为三类：**(1)** 数据集推荐系统。DataLinking [3]和 DataFinder [47]。DataLinking 使用概念频率和词频-逆文档频率提取用户查询与数据集描述的相似性特征。DataFinder 提出了一种基于文本相似度的数据集推荐模型，并使用 BERT 作为数据集描述的嵌入模型。**(2)** 对话推荐系统。DeepSeek-V3:671b [26]（2025年-03- 24）和 Qwen3:235B [55]。**(3)** 基于智能体的对话推荐系统。CoSearchAgent [13] 和 InteRecAgent [17]。CoSearchAgent 是一个多智能体协作系统，能够有效支持多用户对话。

评估指标. 遵循以往研究 [3, 47], , 我们采用常用的推荐指标作为离线评估指标，即 Top-K 召回率、归一化折损累计增益（NDCG）和平均倒数排名（MRR）。由于本文聚焦于面向科学场景的精准推荐，因此重点关注上述指标在 @1、@3 和 @5 处的表现。详细的离线指标信息见附录 A.2。我们还采用多轮对话中成功完成推荐所需的平均轮数（AT）作为指标。在计算 AT 时，t 轮内未成功的推荐记为 $t + 1$ 。对于在线性能评估，我们始终将点击率（CTR）作为主要指标。

实现细节。 我们采用 Qwen-Plus [45] (2025- 04-28) 作为系统的核心大语言模型，用于用户意图解析、工具规划以及构建离线对话数据集。它支持最长 126K 词元的输入。ScienceDB AI 的框架使用 Python 和 LangGraph [22] 实现。我们采用分布式 Qdrant [41] 集群作为在线向量数据库。对于基于对话的模型（如 DeepSeek 和通义千问），我们首先使用近似最近邻（ANN）搜索，根据研究人员的查询检索候选数据集（如图 9 所示的工具结果）。这些候选数据集选自 ScienceDB 中超过 1000 万个数据集。随后，我们将检索到的数据集作为上下文提供给对话大语言模型，由其选出最终推荐结果。所有对比基线均采用默认超参数。对于不支持多轮交互的模型（如 DataFinder 和 DataLinking），我们将多轮查询拆解为一系列单轮查询。对于我们的框架，默认将 N 设为 30。

4.2 整体性能

我们首先评估 ScienceDB AI 及其竞争对手在离线多轮对话式推荐中的整体性能。结果如表2所示。我们的主要发现如下：**(1)** 专为数据集推荐设计的现有模型（DataFinder 和 DataLinking）表现不佳。这些模型主要依赖输入查询与数据集描述之间浅层的语义相似度，因此难以理解研究人员复杂且专业化的需求。值得注意的是，DataFinder 由于依赖简单的关键词相似度，表现尤其糟糕。**(2)** 基于智能体的模型优于基于对话的大语言模型，展现了采用智能体结构的有效性。**(3)** 我们提出的 ScenceDB.AI 在所有评估指标上始终优于所有竞争对手，验证了我们框架的有效性。与最强基线 InteRecAgent 相比，ScenceDB.AI 的性能提升超过 20%。尽管 InteRecAgent 增加了额外模块（例如复杂的规划与反思模块），其表现仍不及我们更精简、高效的设计。**(4)** 根据 AT 的结果，我们得出结论：ScenceDB.AI 找到真实答案所需的轮数最少。与 AT 表现最佳的竞争对手通义千问相比，ScenceDB.AI 在 AT@3 和 AT@5 上分别提升了 8% 和 10%。**(5)** 我们发现，大多数基线模型都能从增大 k的值中显著受益。例如，CoSearchAgent 的召回率从 @1 到 @5 提升了 48.4%。相比之下，ScienceDB AI 仅提升了 3.2%，因为它已经取得了较高的召回率

在靠前的位置，这反映出其能够以较高的初始精确率将正确数据集排在前列。

4.3 运行效率

我们将 ScienceDB AI 与其他基于大语言模型和智能体的对话基线方法进行了比较，以评估其运行效率。图5报告了每个离线对话测试样本的平均推理时间。如图所示，采用自反思模块的 InteRecAgent 推理时间显著高于所有其他模型（518 秒）。尽管 In- teRecAgent 是单智能体模型，其运行速度却慢于基于多智能体的 CoSearchAgent，这凸显了自反思带来的计算成本。令人意外的是，DataLinking 虽然基于简单的关键词相似度，而非大语言模型，但其推理时间仍长于若干基于大语言模型的方法，表明其实现存在低效之处。相比之下，ScienceDB AI 在推理效率和有效性方面均表现出色，因此非常适合部署于现实世界的大规模数据共享服务中。

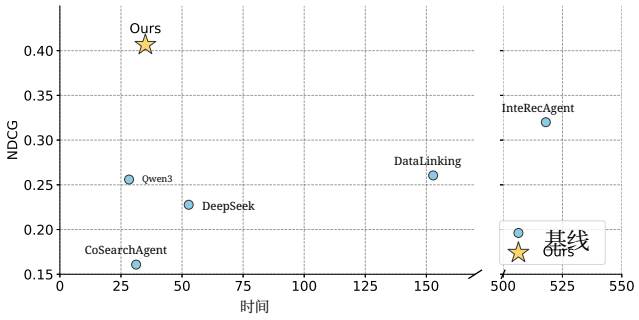


图5：每个测试样本的平均运行时间。

4.4 案例研究

为有效比较性能，我们在图6中展示了案例研究。对于给定的实验级查询，我们比较了两个基于智能体的推荐系统（即 InteRecAgent 和 CoSearchAgent）与 ScienceDB AI 的输出。输入查询见附录中的图9。具体而言，研究人员请求查找有关向熔融铅铋合金注水期间压力升高动力学的数据集。该请求涵盖共晶合金（44.5% Pb–55.5% Bi）、非共晶成分和纯铋。用户还明确要求提供同步诊断输出和分层热工况。

InteRecAgent 和 CoSearchAgent 均返回了 PMCI 数据集。该数据集包含共晶铅铋合金实验的压力和温度测量数据。然而，它未能满足若干关键要求：数据仅涵盖共晶成分，缺少非共晶和纯金属案例的数据。除语义不匹配外，CoSearch 还存在结构性错误。例如，它错误标注数据集枚举编号，导致数据集 ID 与相应描述不匹配。相比之下，我们的 ScienceDB AI 正确识别出了更合适的数据集。该数据集包含 2020年至2022年间开展的纯铅实验所得的高分辨率压力轨迹、同步声学和视频诊断，以及附有完整文档的全面元数据。

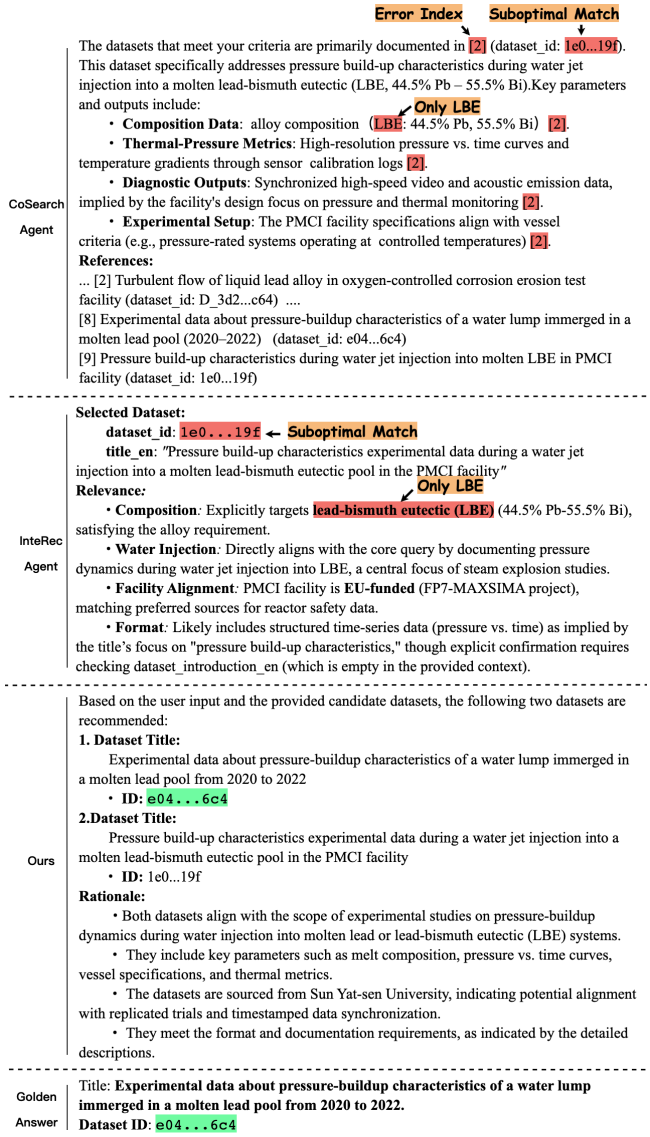


图6：ScienceDB AI及其两个具有竞争力的基于智能体的推荐系统案例研究。

4.5 在线A/B测试

ScienceDB AI 为原有的 ScienceDB 平台引入了全新的搜索界面。我们重点比较 Top-4 排名位置上 ScienceDB AI 与其竞争基线方法的点击率，并以原有的在线基于关键词的搜索页面（https://scidb.cn/en/list?searchList）作为对照。基线系统包含四种检索配置：（1）基于相关性的模糊匹配（相关性/模糊），（2）基于相关性的精确匹配（相关性/精确），（3）基于下载频率的模糊匹配（深度学习/模糊），以及（4）基于下载频率的精确匹配（深度学习/精确）。如图7所示，所有数值均表示我们的模型和可比较的基线方法相对于基于关键词的搜索系统的提升幅度，以百分比表示。我们得出以下发现：**(1)**ScienceDB AI 的点击率显著更高，优于所有基线方法

在所有设置下均如此。这一结论与表2中的离线实验结果一致。值得注意的是，在精确匹配条件下，提升更为显著。**(2)** 基于相关性和基于深度学习的匹配，其性能提升在模糊和精确设置下没有显著差异。结果表明，传统的基于关键词的数据集搜索方法无法捕捉输入查询的语义。相反，这些方法依赖字符串匹配：无论是要求完全一致的精确匹配，还是基于字符相似度的模糊匹配，都无法理解研究人员的科研意图。

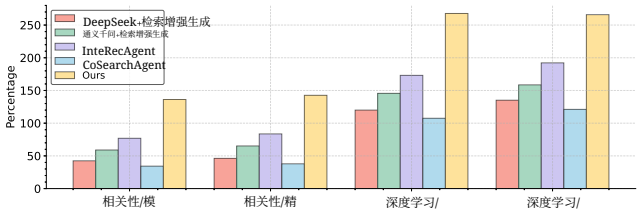


图7：我们的 ScienceDB AI 在在线A/B测试中的性能优于原始检索系统。

5 结论

本文介绍了 ScienceDB AI，这是一个面向大规模科学数据共享的智能体推荐系统，构建于包含超过 1000 万个高质量科学数据集的存储库之上。该系统引入了多项创新：实验意图感知器，用于从复杂查询中提取结构化实验要素；结构化记忆压缩器，用于有效管理多轮对话；以及可信检索增强生成（可信 RAG）框架。可信 RAG 采用两阶段检索机制，并通过可引用科学任务记录（CSTR）标识符提供可引用的数据集引用，从而提高推荐可信度和可复现性。通过使用大规模真实世界数据集开展广泛的离线和在线实验，ScienceDB AI 展现出显著成效：与先进基线方法相比，离线指标提升约 30%；与基于关键词的搜索引擎相比，点击率提升超过 200%。据我们所知，ScienceDB AI 是首个专为大规模科学数据集共享服务量身打造的由大语言模型驱动的对话式推荐系统。

参考文献

[1] Basmah Altaf、Uchenna Akujuobi、Lu Yu 和 Xiangliang Zhang。2019年。基于变分图自编码器的数据集推荐。载于*2019*年*IEEE*国际数据挖掘会议（*ICDM*）。IEEE，11–20。[2] Frank Bell、Raj Chirumamilla、Bhaskar B Joshi、Bjorn Lindstrom、Ruchi Soni 和 Sameer Videkar。2021年。数据共享、数据交换和 Snowflake 数据市场。载于《*Snowflake Essentials: Getting Started with Big Data in the Cloud*》。Springer，299–328。[3] Mohamed Ben Ellefi、Zohra Bellahsene、Stefan Dietze 和 Konstantin Todorov。2016年。用于数据链接的数据集推荐：一种内涵式方法。载于欧洲语义网会议。Springer，36–51。[4] Frédérique Bordignon。2024年。OpenAlex 是一场变革，还是文献计量学/文献计量学家的挑战？载于*Groupe de travail Science Ouverte UDICE*。[5] Dan Brickley、Matthew Burgess 和 Natasha Noy。2019年。Google Dataset Search：在开放网络生态系统中构建数据集搜索引擎。载于万维网大会。1365–1375。

[6] Gustau Camps-Valls、Miguel-Ángel Fernández-Torres、Kai-Hendrik Cohrs、Adrian Höhl、Andrea Castelletti、Aytac Pacal、Claire Robin、Francesco Martinuzzi、Ioannis Papoutsis、Ioannis Prapas等。2025年。用于极端天气和气候事件建模与理解的人工智能。《Nature Communications》16，1（2025年），1919。[7] Xiaomeng Chai、Min Zhang和Hua Tian。2024年。科学智能：百度飞桨的实践。在2024年波特兰国际工程与技术管理会议（PICMET）论文集。IEEE，1–12。[8] Guoxuan Chen、Lianghao Xia和Chao Huang。2025年。LightGNN：用于推荐的简单图神经网络。在第十八届ACM国际网络搜索与数据挖掘大会论文集。549–558。[9] Yuhong Dai、Jianxun Lian、Yitian Huang、Wei Zhang、Mingyang Zhou、Mingqi Wu、Xing Xie和Hao Liao。2025年。利用基于嵌入的记忆为大语言模型预训练上下文压缩器。在计算语言学协会第63届年会论文集（第1卷：长篇论文）。28715–28732。[10] Jiabao Fang、Shen Gao、Pengjie Ren、Xiuying Chen、Suzan Verberne和Zhaochun Ren。2024年。多智能体对话式推荐系统。arXiv预印本 arXiv:2402.01135（2024年）。[11] Sebastian Farquhar、Jannik Kossen、Lorenz Kuhn和Yarin Gal。2024年。利用语义熵检测大语言模型中的幻觉。《Nature》630，8017（2024年），625–630。[12] Lewis Y Geer、Aron Marchler-Bauer、Renata C Geer、Lianyi Han、Jane He、Siqian He、Chunlei Liu、Wenyao Shi和Stephen H Bryant。2010年。NCBI生物系统数据库。《Nucleic Acids Research》38，增刊_1（2010年），D492–D496。[13] Peiyuan Gong、Jiamian Li和Jiaxin Mao。2024年。CoSearchAgent：基于大语言模型的轻量级协作式搜索智能体。在第47届国际ACM信息检索研究与发展大会论文集。2729–2733。[14] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi等。2025年。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。arXiv预印本 arXiv:2501.12948（2025年）。[15] Taicheng Guo、Chaochun Liu、Hai Wang、Varun Mannam、Fang Wang、Xin Chen、Xiangliang Zhang和Chandan K Reddy。2024年。用于推荐的知识图谱增强型语言智能体。arXiv预印本 arXiv:2410.19627（2024年）。[16] Lin He和Vinita Nahar。2016年。学术出版物中的科学数据复用：对Dryad数字知识库的调查。《Aslib Journal of Information Management》68，4（2016年），478–494。[17] Xu Huang、Jianxun Lian、Yuxuan Lei、Jing Yao、Defu Lian和Xing Xie。2025年。推荐智能体：整合大语言模型实现交互式推荐。《ACM Transactions on Information Systems》43，4（2025年），1–33。[18] Shashank Mohan Jain。2022年。Hugging Face。在《NLP Transformers入门：使用Hugging Face库和模型解决问题》中。Springer，51–67。[19] Wei Ju、Zheng Fang、Yiyang Gu、Zequn Liu、Qingqing Long、Ziyue Qiao、Yifang Qin、Jianhao Shen、Fang Sun、Zhiping Xiao等。2024年。深度图表示学习综述。《Neural Networks》173（2024年），106207。[20] Omar Khattab和Matei Zaharia。2020年。ColBERT：通过上下文化后期交互实现高效且有效的段落检索。在第43届国际ACM信息检索研究与发展大会论文集。39–48。[21] Moritz UG Kraemer、Joseph L-H Tsui、Serina Y Chang、Spyros Lytras、Mark P Khurana、Samantha Vanderslott、Sumali Bajaj、Neil Scheidwasser、Jacob Liam Curran-Sebastian、Elizaveta Semenova等。2025年。用于传染病流行建模的人工智能。《Nature》638，8051（2025年），623–635。[22] LangGraph团队。2025年。LangGraph：平衡智能体控制与自主性。https://www.langchain.com/langgraph。[23] Chuang Li、Yang Deng、Hengchang Hu、Min-Yen Kan和Haizhou Li。2025年。ChatCRS：为基于大语言模型的对话式推荐系统融入外部知识与目标引导。在计算语言学协会论文集：北美计算语言学协会2025。295–312。[24] Chengzan Li、Yuanchun Zhou、Xiaohuan Zheng、Zeyu Zhang、Lulu Jiang、Zongwen Li、Pengyao Wang、Jianhui Li、Songyuan Xu和Zhanjie Wang。2022年。追溯中国开放研究数据的发展足迹。《Learned Publishing》35，1（2022年），46–55。[25] Tingting Liang、Chenxin Jin、Lingzhi Wang、Wenqi Fan、Congying Xia、Kai Chen和Yuyu Yin。2024年。LLM-REDIAL：利用大语言模型从用户行为创建的大规模对话式推荐系统数据集。在计算语言学协会论文集：ACL 2024。8926–8939。[26] Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan等。2024年。DeepSeek-V3技术报告。arXiv预印本 arXiv:2412.19437（2024年）。[27] Yuxi Liu、Lianghao Xia和Chao Huang。2024年。SelfGNN：用于序列推荐的自监督图神经网络。在第47届国际ACM信息检索研究与发展大会论文集。1609–1618。[28] Qingqing Long、Haixin Wang、Jinan Sun、Wei Xiang、Yijia Xiao、Yusheng Zhao和Xiao Luo。2025年。针对噪声学习具有鲁棒性的二进制描述符，用于

高效图像检索。在第48届国际ACM信息检索研究与发展大会论文集。2695–2699。[29] Qingqing Long、Lingjun Xu、Zheng Fang和Guojie Song。2021年。HGK-GNN：基于异构图核的图神经网络。在第27届ACM知识发现与数据挖掘大会论文集。1129–1138。[30] Donald Loveland、Xinyi Wu、Tong Zhao、Danai Koutra、Neil Shah和Mingxuan Ju。2025年。从矩阵秩视角理解并扩展协同过滤优化。在2025年ACM Web会议论文集。436–449。[31] Junyu Luo、Weizhi Zhang、Ye Yuan、Yusheng Zhao、Junwei Yang、Yiyang Gu、Bohan Wu、Binqi Chen、Ziyue Qiao等。2025年。大语言模型智能体：方法论、应用与挑战综述。arXiv预印本 arXiv:2503.21460（2025年）。[32] Ron L'Esteve。2022年。Databricks。在《Azure数据湖仓工具包：使用Delta Lake、Apache Spark、Databricks、Synapse Analytics和Snowflake在Azure上构建并扩展数据湖仓》中。Springer，83–139。[33] D-Lib杂志。2011年。Dataverse网络®：用于共享、发现和保存数据的开源应用。《D-Lib杂志》17，1/2（2011年）。[34] Eric Martin和Daniel N Osherson。1998年。科学探究要素。麻省理工学院出版社。[35] Marcos Martinez-Romero、Matthew Horridge、Nilesh Mistry、Aubrie Weyhmiller、Jimmy K Yu、Alissa Fujimoto、Aria Henry、Martin J O' Connor、Ashley Sier、Stephanie Suber等。2025年。RADx 数据中心：面向公平、协调一致的新冠肺炎数据的云平台。arXiv预印本 arXiv:2502.00265（2025年）。[36] William F McComas。1998年。科学本质的主要要素：消除误解。在《科学教育中的科学本质：理论依据与策略》中。Springer，53–70。[37] Lisa Messeri和Molly J Crockett。2024年。科学研究中的人工智能与理解幻觉。《Nature》627，8002（2024年），49–58。[38] Anton Ninkov、Kathleen Gregory、Isabella Peters和Stefanie Haustein。2021年。DataCite上的数据集——初步文献计量学调查。在第18届国际科学计量学与信息计量学大会（ISSI 2021）论文集。1–6。[39] Chunjong Park、Anas Awadalla、Tadayoshi Kohno和Shwetak Patel。2021年。利用数据集偏移检测实现可靠且可信的健康机器学习。《神经信息处理系统进展》34（2021年），3043–3056。[40] Ken Peffers、Tuure Tuunanen、Marcus A Rothenberger和Samir Chatterjee。2007年。信息系统研究的设计科学研究方法论。《Journal of Management Information Systems》24，3（2007年），45–77。[41] Qdrant团队。2024年。Qdrant：向量相似性搜索引擎。https://github.com/qdrant/qdrant。[42] Najla Rettberg和Birgit Schmidt。2012年。OpenAIRE——为欧洲研究人员构建协作式开放获取基础设施。《LIBER季刊：欧洲研究图书馆协会杂志》22，3（2012年），160–175。[43] Miguel-Angel Sicilia、Elena García-Barriocanal和Salvador Sánchez-Alonso。2017年。开放数据集存储库中的社区策展：来自Zenodo的洞见。《Procedia Computer Science》106（2017年），54–60。[44] Ying Sun、Hengshu Zhu、Lu Wang、Le Zhang和Hui Xiong。2024年。大规模在线求职行为揭示新冠肺炎期间的劳动力市场变化。《自然·城市》1，2（2024年），150–163。[45] Qwen团队。2024年。Qwen2技术报告。arXiv预印本 arXiv:2407.10671（2024年）。[46] Mike Thelwall和Kayvan Kousha。2016年。FigShare：学术资源共享的通用存储库？《Online Information Review》40，3（2016年），333–346。[47] Vijay Viswanathan、Luyu Gao、Tongshuang Wu、Pengfei Liu和Graham Neubig。2023年。DataFinder：基于自然语言描述的科学数据集推荐。在计算语言学协会第61届年会论文集（第1卷：长篇论文）。计算语言学协会，10288–10303。[48] Wei Wang、Kepan Chen、Ting Jiang、Yiyang Wu、Zheng Wu、Hang Ying、Hang Yu、Jing Lu、Jinzhong Lin和Defang Ouyang。2024年。由人工智能驱动的可电离脂质合理设计，用于mRNA递送。《Nature Communications》15，1（2024年），10804。[49] Yancheng Wang、Ziyan Jiang、Zheng Chen、Fan Yang、Yingxue Zhou、Eunah Cho、Xing Fan、Yanbin Lu、Xiaojiang Huang和Yingzhen Yang。2024年。RecMind：由大语言模型驱动的推荐智能体。在计算语言学协会论文集：北美计算语言学协会2024。4351–4364。[50] Zhefan Wang、Yuanqing Yu、Wendi Zheng、Weizhi Ma和Min Zhang。2024年。MACRec：用于推荐的多智能体协作框架。在第47届国际ACM信息检索研究与发展大会论文集。2760–2764。[51] Chuyuan Wei、Ke Duan、Shengda Zhuo、Hongchun Wang、Shuqiang Huang和Jie Liu。2025年。利用检索增强大语言模型增强推荐系统。《人工智能研究杂志》82（2025年），1147–1173。[52] Joss Winn。2013年。开放数据与学术界：对CKAN用于科研数据管理的评估。（2013年）。[53] Wujiang Xu、Yunxiao Shi、Zujie Liang、Xuying Ning、Kai Mei、Kun Wang、Xi Zhu、Min Xu和Yongfeng Zhang。2025年。Instructagent：通过大语言模型智能体构建用户可控推荐系统。arXiv电子预印本（2025年），arXiv–2502。[54] Yuwei Yan、Yu Shang、Qingbin Zeng、Yu Li、Keyu Zhao、Zhiheng Zheng、Xuefei Ning、Tianji Wu、Shengen Yan、Yu Wang 等。2025年。AgentSociety挑战赛：为网络平台上的用户建模与推荐设计大语言模型智能体。载于*ACM万维网大会2025*年伴随论文集。2963–2967。[55] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。2025年。Qwen3技术报告。*arXiv*预印本 *arXiv:2505.09388*（2025年）。[56] Junjie Zhang、Yupeng Hou、Ruobing Xie、Wenqi Sun、Julian McAuley、Wayne Xin Zhao、Leyu Lin 和 Ji-Rong Wen。2024年。AgentCF：利用自主语言智能体进行协同学习的推荐系统。载于*ACM万维网大会2024*论文集。3679–3689。[57] Yu Zhang、Yiwen Zhang、Yi Zhang、Lei Sang 和 Yun Yang。2025年。揭示对比学习在协同过滤邻域聚合方面的能力。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议论文集。1985–1994。[58] Yuyue Zhao、Jiancan Wu、Xiang Wang、Wei Tang、Dingxian Wang 和 Maarten De Rijke。2024年。让我来为你完成：通过工具学习实现大语言模型赋能的推荐。载于第*47*届国际*ACM SIGIR*信息检索研究与发展会议论文集。1796–1806。[59] Guangfeng Zhou、Domnita-Valeria Rusnac、Hahnbeom Park、Daniele Canzani、Hai Minh Nguyen、Lance Stewart、Matthew F Bush、Phuong Tran Nguyen、Heike Wulff、Vladimir Yarov-Yarovoy 等。2024年。一种加速药物发现的人工智能虚拟筛选平台。《*Nature Communications*》15卷，1期（2024年），7761。[60] Yuanchun Zhou、Pengyao Wang、Chengzan Li、Zongwen Li、Lulu Jiang、Zeyu Zhang 和 Jia Liu。2024年。科学数据银行（ScienceDB）的可信系统与国际服务能力建设。载于《中国电子科学蓝皮书*2023*》。Springer，427–445。[61] Zhihong Zhu、Fan Zhang、Yunyan Zhang、Jinghan Sun、Zhiqi Huang、Qingqing Long、Bowen Xing 和 Xian Wu。2025年。多模态意图识别综述：最新进展与新前沿。计算语言学协会*EMNLP 2025*年论文集：研究成果（2025年），15223–15236。

A 附录

A.1 详细离线数据集信息

源数据。图8展示了一个具有代表性的数据集条目及其典型的结构化元数据，包括标题、作者信息、分类体系分类、关键词和文本描述。这类条目捕捉了用于索引和检索的关键信息，并为数据集推荐和语义理解等下游任务奠定基础。

一个候选数据集样本。

"标题": "关于2020年至2022年熔融铅池中水团浸入后的压力buildup特性",

"cstr": "31253.11.sciencedb.j00186.00022", **"dataSetPublishDate"**: "2023-02-24T06:52:19Z",**"author"**: [{"name": "...", "organizations": ["..."] }, {"name": "...", "organizations": ["..."]}],**"taxonomy"**: ["code": "490","nameZh": "","nameEn": "核科学与技术"], **"keywordEn"**: ["铅冷快堆","蒸汽发生器传热管破裂事故","压力buildup特性","实验研究"], **"introduction"**: "为了解熔融铅池中水滴浸入后的压力buildup特性——这是铅冷快堆发生蒸汽发生器传热管破裂事故时的关键现象——中山大学于2020年至2022年开展了多项实验，将水团注入熔融铅池。为深入了解熔体材料的影响，本研究将铅实验与文献中的铅铋共晶（LBE）实验进行了比较。实验采用的参数包括水体积、水的形状、水过冷度、熔池深度和熔体温度。发生CCI的相互作用容器为不锈钢圆柱容器，内径为250 mm，高750 mm，设计压力为40 MPa。相互作用容器壁上安装了多个传感器，用于获取熔池和覆盖气体的温度及压力变化趋势。"

图 8：候选数据集样本，包含元数据和描述。

离线评估数据集构建伪代码。 算法 2 概述了根据用户的历史交互构建模拟多轮对话条目 e的流程。给定用户 ID u、历史项目序列 $H = [h_1, h_2, ..., h_n]$、选定的目标索引 i、模板模块 T以及最大交互轮数 R，该算法会生成一段合成对话，反映真实且具有挑战性的信息搜寻过程。在**步骤 1**中，算法选取一个固定长度的历史窗口 $H_{\text{sel}} = H[i - L: i]$位于目标索引 i之前。目标项目$d = H[i]$代表用户真正想要检索的数据集。使用 u、i、T和 H_{sel}初始化一个新的对话条目 e。将 H_{sel} 中各项目的标题拼接成字符串 s，并传入传递给系统提示生成器 T.sys_prompt(s)，并附加到条目中。在步骤 2 中，GenFakeRequest 根据 T、s 和真实数据集 d，生成伪装用户查询 q_0 及其对应的监督掩码 mask。该查询旨在间接表达用户的真实意图。掩码存储在 e 中，查询则使用 T.fake_request(q_0) 格式化，然后作为用户消息附加。步骤 3 模拟多轮对话循环。在每一轮 r 中，都会从 e 提取最新的用户消息和助手消息（u_r、a_r），并使用 T.tool_query(s、a_r、u_r) 构建新的工具查询 q_r。如果 q_r 包含预定义的任务结束标记，循环便会终止。否则，查询会通过 Search 发送至检索后端，返回的文档会格式化为 R_r。随后，助手生成回复 $a_{r+1} = T$.generate_response(s、a_r、u_r、R_r)，用户接着发送后续消息 $u_{r+1} = T$.user_followup(s、d、a_{r+1})。这两条消息都会附加到 e 中；如果 u_{r+1} 也包含结束标记，循环则会提前停止。最后，在步骤 4 中，通过 T.truth_response(d) 生成与目标数据集 d 对应的真实答案，并将其作为助手的最终回复附加到对话中。

算法 2：生成多轮对话

输入：用户 ID u、历史记录 $H = [h_1, h_2, ..., h_n]$、目标索引 i、模板 T、最大轮数 R
输出：对话条目 e

// **步骤 1：初始化条目**
$H_{\text{sel}} \leftarrow H[i - L: i]$;
$d \leftarrow H[i]$; // 目标数据集
$e \leftarrow$ InitEntry(u, i, T, H_{sel});
$s \leftarrow$JoinTitles(H_{sel});
将（系统提示词 T.sys_prompt(s)）追加到 e;
// **步骤 2：生成初始用户查询**
(q_0，掩码) $\leftarrow$GenFakeRequest(T，s，d);
e.掩码 $\leftarrow$掩码;
将（用户请求 T.虚假_请求(q_0)）追加到 e;
// **步骤 3：多轮模拟**
对于 $r = 1$ **到** $R - 1$ **执行**
  (u_r，a_r) $\leftarrow$最后消息(e);
  $q_r \leftarrow T$.工具_查询(s，a_r，u_r);
  将（调用工具 q_r）追加到 e;
  如果结束标记(q_r) **则**
    跳出;
  $R_r \leftarrow$ Search(q_r);
  将（工具结果，JSON(R_r)）追加到 e;
  $a_{r+1} \leftarrow T$.生成_响应(s，a_r，u_r，R_r);
  将（助手响应 a_{r+1}）追加到 e;
  $u_{r+1} \leftarrow T$.用户_后续请求(s，d，a_{r+1});
  将（用户请求 u_{r+1}）追加到 e;
  如果结束标记(u_{r+1})**则**
    跳出;
// **步骤 4：最终标准答案**
将（最终答案 T.真实答案_响应（d））追加到 e;
返回 e;

离线模拟对话式数据集推荐示例。

{
“操作_类型”：**“模拟用户请求”**，
“内容”: ”
数据集范围：记录熔融铅铋合金注水过程中压力升高动力学的实验数据集（包括共晶铅铋合金：44.5% 铅-55.5% 铋，非共晶配比如 20% 铅-80% 铋或 80% 铅-20% 铋）或纯铋。研究必须报告在分层（例如底部 400°C、顶部 500°C）和均匀熔体温度（~500°C）两种条件下的结果。关键参数：
1.**成分数据**：完整的合金成分（铅/铋的质量百分比）、熔体材料性质（密度、热导率、黏度）。
2. **热压指标**：高分辨率压力-时间曲线、具有空间分辨率的温度梯度（例如轴向热电偶）、注水参数（50 mL 体积、80°C 过冷水、液滴形态）。
3. **诊断输出**：同步高速视频（破碎模式、蒸汽层塌缩）和声发射（频率频谱、与压力尖峰相关的振幅突发），并包含将前兆信号（如蒸汽塌陷）与压力动力学关联起来的元数据。
4. **实验设置**：容器规格（如不锈钢材质，内径 250 mm × 750 mm 高，耐压 40 MPa）、传感器校准日志用于压力/温度测量、熔体制备与注入方案（感应加热、热电偶阵列）。排除项：- 非金属熔体或非水冷却剂（如钠、二氧化碳）。优先来源：- OECD/NEA 数据库、国际原子能机构反应堆安全计划、FP7-MAXSIMA 数据集，或其他由欧盟/美国核管理委员会资助的蒸汽爆炸物理学实验。- 重复试验（每种条件 ≥5 次），时间戳需与压力、热学和诊断数据对齐。- 纯铅与富铋合金的对比数据集。格式要求：- 结构化时间序列数据（CSV、HDF5、MATLAB .mat、Excel）。- 带时间戳的视频（AVI、MP4）及其元数据。- 文档：README 文件、传感器校准记录、实验示意图。搜索关键词：熔融金属与水的相互作用、蒸汽爆炸压力动力学、铅铋合金热分层、铋冷却剂安全实验、蒸汽爆炸中的声学前兆、反应堆冷却系统碎裂研究。目的：对峰值压力、压力上升速率和声学前兆进行基准测试；根据先前的实验数据验证诊断同步（例如视频与声学信号对齐）。实验数据。”
},
{
"action_type"：**"工具结果"**,
"内容"：[
[0.7737，**数据集_id: e04...6c4**，标题：关于熔融铅中水团的压力 buildup 特性的实验数据池中2020年至2022年。，数据集_引言：...],
[0.7697，**数据集_id: 1e0...19f**，标题：水射流注入熔融PMCI设施中铅铋共晶池的压力 buildup 特性实验数据，数据集_引言：...],
...,
[0.6674，**数据集_id: D_3d2...c64**，标题：氧控腐蚀侵蚀试验设施中的液态铅合金湍流，数据集_引言：...],
..."
},
{
"action_type"：**"模型响应"**,
"内容"：「.....」
},
{
"action_type"：**"真实答案"**,
"内容"：「根据您的请求，以下是您想要的数据集：关于水的压力 buildup 特性的实验数据团在熔融铅池中浸没的情况，时间范围为2020年至2022年。数据集 ID 为：e04...6c4。」
}

图9：离线对话式数据集推荐的模拟示例。

具体数据集实例/样本。 我们提供了一个具体的离线数据集实例，如图 9 所示。每个数据集样本由研究人员的模拟查询、工具结果、模型输出的回复以及真实答案组成。我们通过对详细的实验需求进行建模来构建模拟用户查询，例如成分数据、热压指标、诊断输出和实验设置。与简单的关键词查询相比，这种方法能够更全面地捕捉真实的用户需求，从而更好地评估数据集推荐系统。基于这一请求，系统从一个结构化的数据库将其记录为工具结果（候选数据集）。最后，真实答案提供了模拟用户实际想要查找的数据集。

A.2 评估指标详细信息

为评估数据集推荐模型的性能，我们采用以下广泛使用且具有代表性的基于排名的指标：Recall@k、NDCG@k 和 MRR@k，其中 $k \in \{1、3、5\}$。这些指标定义如下：

Recall@k。Recall@k 用于衡量真实相关项目是否排在前k位以内：

$$\text{Recall@}k = \begin{cases} 1, & \text{if relevant item is ranked } \le k. \\ 0, & \text{otherwise.} \end{cases}$$

在我们的设定中，每个查询只有一个相关数据集，因此 Recall@k 用于评估位置 k处的命中率。

NDCG@k（归一化折损累计增益）。NDCG@k 会考虑相关项在排序列表中的位置，并为排名靠前的项赋予更高权重。其定义如下：

$$\text{NDCG@}k = \frac{1}{\log_2(r+1)} \quad \text{if relevant item is at rank } r \le k$$

否则，NDCG@k $= 0$ 当只有一个相关项时，理想折损累计增益（IDCG@k）为 1，因此 NDCG@k 可简化为单个位置折损。

MRR@k（平均倒数排名）。MRR@k 衡量第一个相关项排名位置的倒数，并在 k处截断：

$$\text{MRR@}k = \begin{cases} \frac{1}{r}, & \text{if relevant item is at rank } r \le k. \\ 0, & \text{otherwise.} \end{cases}$$

我们报告所有查询的平均 MRR@k。所有指标均在测试集上取平均，并在 $k = 1$、3、5 处进行评估，以衡量不同深度下的排序质量。

为进一步评估多轮对话中的推荐效率，我们提出一项新指标：**平均轮数（平均轮次）**。该指标衡量模型能够多早推荐出对话中的正确数据集。具体而言，对于每段多轮对话，我们找出模型的回答首次在前k 个结果中包含真实数据集的轮次 t。该对话的平均轮数得分定义为：

$$\text{AT} = \begin{cases} t, & \text{if the correct dataset appears in turn } t \le T \\ T+2, & \text{if the correct dataset is not found in any turn} \end{cases}$$

其中，T 表示对话轮次总数。$T + 2$ 的惩罚机制确保对模型完全未能成功的对话给予适当惩罚。最终的平均轮数得分通过对所有对话取平均计算得出。

多轮对话的指标扩展。与传统的单轮场景不同，我们的数据集包含多轮对话式查询，用户会逐步细化其请求。为反映这一过程，我们采用全局 top-k 评估策略：不再对各轮次的指标（Recall@k、NDCG@k、MRR@k）取平均，而是按逆时间顺序（从最后一轮到第一轮）串联模型响应，并在生成的排序列表上计算指标，优先考虑较晚出现、更加具体的意图。

然而，标准指标无法区分正确数据集是在较早还是较晚的轮次被识别出来。为衡量交互效率，我们提出平均轮数（AT），用于反映推荐正确数据集所需的最早轮次。AT 越低，表示任务解决越快、理解越准确。将 AT 与标准指标结合，可更全面地评估推荐质量和效率。