

核物理 AI 研究助手与 arXiv 向量数据库

庞龙刚^{1,2}

1(华中师范大学 夸克与轻子教育部重点实验室 武汉 430079)

2(华中师范大学 人工智能与计算物理研究中心 武汉 430079)

摘要 面对跨学科科学文献的指数级增长与现有检索系统的局限性,本研究基于 arXiv 平台 266 万篇论文数据集,创新开发了融合向量语义检索与大型语言模型(Large Language Model, LLM)分析的智能系统。通过构建论文向量数据库实现语义相似性初筛,结合 LLM 上下文推理优化排序,有效解决了传统关键词搜索的语义鸿沟问题以及 LLM 的幻觉问题。在核物理领域的应用表明,该系统能精准定位跨学科解决方案,对比特定任务上的关键词检索和向量相似度检索,前 10 篇文献的查全率从 10% 跃升到 60%,查准率从 20% 跃升到 90%。项目开源提供三大核心模块:1) 全量论文向量数据库;2) 智能检索优化框架(含查询生成、相关性分析等智能体);3) PDF 深度解析工具链。本研究突破性地将语义检索与 LLM 推理相结合,为应对知识爆炸时代的科研挑战提供了可扩展的解决方案(开源地址: https://gitee.com/lgpang/arxiv_vectordb)。

关键词 arXiv 向量数据库, 大语言模型智能体, 深度求索, AI 科学家

中图分类号 O571, TP391.1

DOI: 10.11889/j.0253-3219.2025.hjs.48.250108

CSTR: 32193.14.hjs.CN31-1342/TL.2025.48.250108

Nuclear physics AI research assistant and arXiv vector database

PANG Longgang^{1,2}

1(Key Laboratory of Quark and Lepton Physics (MOE) & Institute of Particle Physics, Central China Normal University, Wuhan 430079, China)

2(Artificial Intelligence and Computational Physics Research Center, Central China Normal University, Wuhan 430079, China)

Abstract [Background] The exponential growth of scientific literature, particularly in physics and nuclear physics, poses significant challenges for researchers to track advancements and identify cross-disciplinary solutions. While large language models (LLMs) offer potential for intelligent retrieval, their reliability is hindered by inaccuracies and hallucinations. The arXiv dataset (2.66 million papers) provides an unprecedented resource to address these challenges. **[Purpose]** This study aims to develop a hybrid retrieval system integrating vector-based semantic search with LLM-driven contextual analysis to enhance the accuracy and accessibility of scientific knowledge across disciplines. **[Methods]** We processed 2.66 million arXiv paper titles/abstracts using BGE-M3 model to generate 1 024-dimensional vector representations. Cosine similarity metrics were computed between user queries (vectorized via the same model) and pre-encoded paper vectors for preliminary semantic ranking. The top 50 candidates underwent contextual relevance analysis by DeepSeek-r1, which evaluated technical depth, methodological alignment, and cross-domain connections through multi-step reasoning. A nuclear physics case study validated the

国家自然科学基金(No.12075098, No.12435009, No.CCNU24JC011)、华中师范大学中央高校基本科研业务费项目资助

第一作者: 庞龙刚, 男, 1984 年出生, 2012 年于中国科学技术大学获博士学位, 研究领域为高能核物理、计算物理与人工智能, E-mail: lgpang@ccnu.edu.cn

收稿日期: 2025-03-12, 修回日期: 2025-04-18

Supported by National Natural Science Foundation of China (No.12075098, No.12435009, No.CCNU24JC011), the Fundamental Research Funds for the Central Universities at Central China Normal University

First author: PANG Longgang, male, born in 1984, graduated from University of Science and Technology of China with a doctoral degree in 2012, focusing on high-energy nuclear physics, computational physics, and artificial intelligence, E-mail: lgpang@ccnu.edu.cn

Received date: 2025-03-12, revised date: 2025-04-18

system using 1 000 AI-human-annotated documents. The framework incorporating four specialized agents: query generation, relevance scoring, structured data correction, and PDF analysis. **[Results]** We constructed a vector database comprising 2.66 million arXiv papers (including titles and abstracts), occupying 30 GB of disk space. Our vector-based semantic search system demonstrated superior performance in a nuclear physics query benchmark, achieving 90% precision and 60% recall for the top-10 retrieved documents. This significantly outperformed traditional keyword-based search methods, which yielded only 20% precision and 10% recall under the same evaluation conditions. **[Conclusions]** By synergizing vector semantics with LLM reasoning, this work establishes a new paradigm for scientific knowledge retrieval that effectively bridges disciplinary divides. The open-sourced system (https://gitee.com/lgpang/arxiv_vectordb) provides researchers with scalable tools to navigate literature complexity, demonstrating particular value in identifying non-obvious interdisciplinary connections.

Key words arXiv vector database, Large language model agent, DeepSeek, AI scientist

人工智能能否彻底改变我们在核物理及其他科学领域开展研究的方式? 随着科学文献的指数级增长, 研究人员在跟上最新进展和识别跨学科联系方面面临着越来越大的挑战。arXiv 等大规模数据集的可用性, 结合 ChatGPT 和 DeepSeek 等大型语言模型 (Large Language Models, LLMs) 的能力, 为解决这些挑战提供了前所未有的机遇。

arXiv 数据集是一个全面的科学预印本存储库, 已成为核物理、材料科学、天体物理学和量子计算等多个领域研究人员的宝贵资源。尽管 arXiv 具有巨大的潜力, 但其庞大的信息量给研究人员带来了显著挑战, 尤其是在识别相关研究、揭示跨学科见解以及紧跟新兴趋势方面。传统的文献综述和知识发现方法耗时且往往无法捕捉不同科学领域之间的微妙关系。近年来, 自然语言处理 (Natural Language Processing, NLP) 和机器学习, 特别是 LLMs 的发展, 为解决这些挑战开辟了新途径。然而, 这些模型仍面临幻觉和检索不准确等局限性, 这阻碍了它们在科学应用中的有效性。

主要挑战在于如何利用 arXiv 数据集^[1]和大语言模型开发智能工具, 以提高科学研究的效率和准确性。当前的方法在整合跨学科知识、构建有意义的知识图谱以及生成可靠的自动化文献综述方面常常遇到困难。此外, 识别未解决的研究问题和预测未来趋势的能力仍然不足。这些局限性凸显了需要创新解决方案来弥合庞大科学数据集与可操作研究见解之间的差距。

最近有两个技术可以帮助我们解决这些问题, 构建 AI 科学家系统。一个是把文本编码为向量, 使用向量余弦相似度进行语义检索的检索增强生成系统 (Retrieval Augmented Generation, RAG), 另一个是大语言模型智能体技术, 赋予 DeepSeek 等 LLM 语言处理之外的能力, 比如记忆能力 (RAG)、使用工具能力、反思能力, 从而让其能够更好地完成提示

词指定的功能。通过向量数据库和多个 LLM 智能体的结合, 有望搭建一个了解 arXiv 上 266 万篇文献, 并进行智能化检索和分析的 AI 科学家系统, 改变核物理^[2-52]以及其他科学领域的研究范式。

许多 AI 科研助手 (如 Elicit、Consensus、Scite.ai) 通过利用 arXiv 论文的元数据 (包括标题、摘要和关键词) 构建向量数据库, 并采用语义相似度搜索 (如余弦相似度) 来实现用户查询匹配。主要技术包括以下三个方面:

1) 嵌入模型: 采用 BERT、SciBERT、BGE-M3 等预训练模型将论文文本编码为高维向量表示;

2) 混合检索: 结合传统关键词检索 (如 BM25 算法) 与语义向量检索, 以提升系统的查全率;

3) 个性化推荐: 基于用户检索历史优化结果排序 (如 Colandr、Semantic Scholar 等系统)。

与此同时, LLM 在学术文献分析领域也展现出显著的应用价值, 具体表现在:

1) 摘要生成: 利用 ChatGPT、DeepSeek 等模型解析 arXiv 论文摘要并生成简明概述 (如 ExplainPaper 工具);

2) 问答系统: 支持用户以自然语言提问 (如 "该研究的创新点是什么?"), 系统从 PDF 全文或摘要中提取相关信息进行回答 (如 SciSpace、Kimi Chat 等平台);

3) 跨学科关联: 通过 LLM 识别不同学科领域论文间的潜在关联 (如 Litmaps、ResearchRabbit 等工具)。

然而, 当前这些系统存在两个主要局限性: 一是多数系统未开源, 二是需要支付较高的查询费用。这在一定程度上限制了科研人员根据自身需求进行定制化研究的可能性。本文旨在探索利用 arXiv 数据集和大语言模型智能体推进核物理及相关领域的若干有前景的研究方向。具体而言, 我们提出开发以下工具和系统: 1) 智能跨学科知识检索系统; 2) 向

量嵌入的知识库;3)缓解 LLM 幻觉的专用模型;4)识别未解决研究问题的系统;5)自动化文献综述生成工具;6)基于时间序列的趋势分析工具;7)个性化研究助手;8)用于科学文献语义理解和推理的高级模型。这些目标旨在提高科学研究的效率、准确性和跨学科潜力。我们初步搭建了 arXiv 向量数据库,实现了向量余弦相似度检索,大语言模型 Agent 相关性分析,解决 LLM 的幻觉问题,生成相关文献的 bibtex 引用列表功能,并将代码开源给领域内的学者使用。使用这个系统,我们检索了深度学习在核物理状态方程方面的研究,并自动化地给出了相关的文献[3-51]。

本文讨论了基于 arXiv 向量数据库的 AI 科学家系统构建时的各个模块。比如 arXiv 数据集的内容简介,基于 BGE-M3^[52]的向量数据库的生成与检索,多个不同功能大语言模型智能体的构建(检索问题生成智能体、文献相关度分析智能体、结构化数据生成与修正智能体、智能 pdf 正文阅读与总结智能体)以及向量数据库与传统数据库的结合。介绍了 AI 科学家系统的安装、使用以及 30 GB 向量数据库的下载,另外介绍了这个系统的未来改进方案,包括未来研究方向和未解决科学问题的识别,领域研究问题和热点的时间序列分析以及最终的自动化文献综述与知识发现。

通过解决这些研究方向,我们旨在为利用 arXiv 和 LLMs 创建更高效、智能和跨学科的科学工具有提供路线图。

1 方法介绍

1.1 arXiv 数据集

arXiv 是由 Paul Ginsparg 于 1991 年创立的学术资源平台,目前由康奈尔大学维护和运营。在过去的 30 年中,arXiv 通过提供开放获取的学术文章服务于公众和研究界。这些文章涵盖了多个学科的研究成果,包括物理学、数学、统计学、计算机科学、电气工程和金融学等领域。这一丰富的资源库提供了广泛的内容,但对于科研工作者以及年轻的学术爱好者来说,也可能导致信息过载。如何从这个巨大的开放数据库中获取新的知识和见解,填补不同学科研究前沿的知识鸿沟,成为非常重要的未解决问题。为了使 arXiv 更易于访问,康奈尔大学与合作者在 Kaggle 上提供了一个免费、开放且机器可读的 arXiv 数据集^[1]。该资源库目前包含 266 万多篇文章,涵盖文章标题、作者、类别、摘要和全文 PDF 等数据。

Kaggle 上的 arXiv 元数据数据集是原始 arXiv 数据的镜像。由于完整数据集规模庞大(1.1 TB 且仍在增长),该数据集仅提供了一个 JSON 格式的元数据文件。文件中包含每篇论文的条目,但不包括完整的 PDF 文件。每个条目包含以下字段:

id: arXiv ID
 submitter: 论文提交者
 authors: 论文作者
 title: 论文标题
 comments: 附加信息,如页数和图表数量
 journal-ref: 论文发表的期刊引用
 abstract: 论文摘要
 categories: arXiv 系统中的类别/标签
 versions: 版本历史

在获取文章的 arXiv ID 后,可以通过以下 URL 直接下载对应的 PDF 文件: <https://arxiv.org/pdf/{id}>。

如图 1 所示,对下载的 arXiv 数据集中不同类别文章进行数量统计,发现在与核物理相关的研究中,核理论方面的文章(nucl-th)58 834 篇,核实验方面的文章(nucl-ex)26 586 篇,高能唯像文章(hep-ph)185 401 篇,高能理论文章(hep-th)171 695 篇。这么庞大的数据集,如果想通过人工语义检索的方式,获得某个研究内容的相关文章,将会非常艰难。构造向量数据库,搭建基于向量余弦相似度和大语言模型相关度分析的智能检索系统,成为现代科学研究的重要需求。

1.2 向量数据库与大语言模型 Agent

下载的 JSON 文件包含超过 266 万个条目,每个条目中的元数据以 JSON 格式存储。我们将每篇文章的标题和摘要拼接为一个字符串,并使用 BGE-M3^[53]嵌入工具将其编码为 1 024 维向量。向量的每个元素为 16 位浮点数(float16)。这创建了一个 1 024 维的语义空间,其中每篇文章的标题和摘要对应于该空间中指向特定方向的向量。生成的向量数据库占据硬盘的空间约 30 GB,其中最新的 arXiv 文章更新到 2025 年 2 月。未来可以将新的文章向量化,补充到此向量数据库。

向量数据库中的检索过程使用余弦相似度来计算向量之间的相关性:

$$\cos \theta = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|} \quad (1)$$

两个长度为 1 的单位向量,满足关系 $|\vec{v}_1| = |\vec{v}_2| = 1$ 。这两个向量的点乘表示了两个向量的相似程度。如果 $\cos \theta = 1$,则两个向量相等,指向高维语

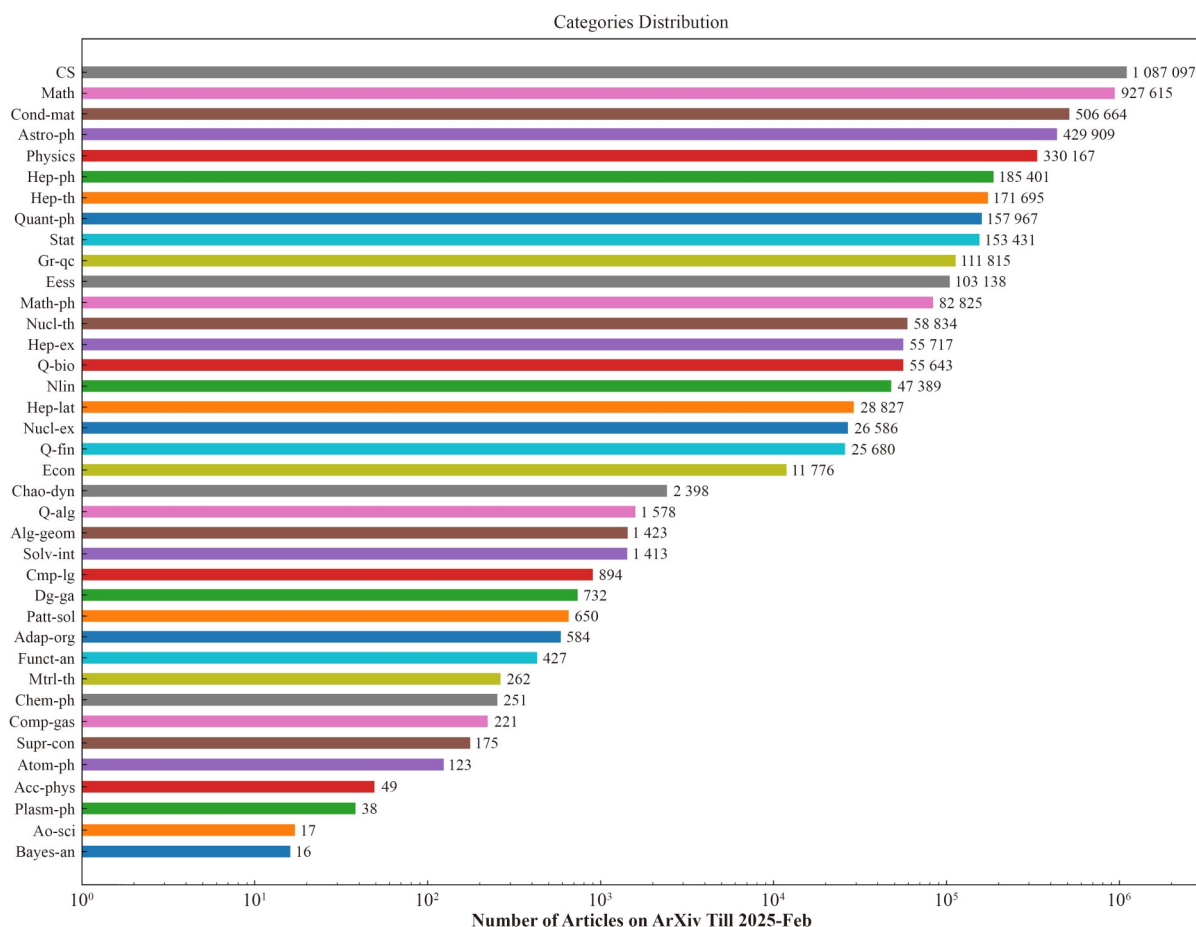


图1 开源的arXiv数据集中不同类别文章数量统计

其中核理论文章(nucl-th)58 834篇,核实验文章(nucl-ex)26 586篇,高能唯像文章(hep-ph)185 401篇,高能理论文章(hep-th)171 695篇

Fig.1 The statistics of article counts across different categories in the open-access arXiv dataset

There are 58 834 articles in nuclear theory (nucl-th), 26 586 articles in nuclear experiment (nucl-ex), 185 401 articles in high-energy phenomenology (hep-ph), and 171 695 articles in high-energy theory (hep-th).

义空间中的同一个方向,表示两个向量的相关程度最高。如果 $\cos \theta = 0$,则两个向量在语义空间正交,表示两个词语、句子或段落的相关性很低。如果 $\cos \theta = -1$,则代表两个向量在语义空间存在反义关系。使用向量检索而不是关键词检索的方法,可以从大量的研究文献中检索出没有共同关键词,但语义非常相近的研究内容。

大语言模型智能体(LLM Agent)被斯坦福大学教授吴恩达称为有可能超越下一代大语言模型基座的技术^[54]。他解释道,传统的大语言模型通常采用单次生成模式(one-shot generation),这种模式类似于要求作者在不允许回溯修改的情况下一性完成散文创作。尽管面临如此严苛的约束条件,大语言模型仍展现出卓越的文本生成能力。可知,如果让大语言模型以人类的工作方式运行,允许它进行任务分解和规划、反思、修改、查阅资料、上网搜索以及调用外部工具(比如计算器、编译器、代码调试器),

则大语言模型的能力将得到极大的提高。这种技术被称为LLM Agent。在使用大语言模型编写程序并让人类打分的能力对比中,GPT-3.5一次成型编写的代码获得48分,GPT-4一次成型的代码获得67分,大语言模型基座GPT-4的能力要明显高于GPT-3.5。但是,给GPT-3.5以特殊能力,让它成为LLM Agent,则可以获得95分的高分,远超新一代的大语言模型基座GPT-4。GPT-3.5的LLM Agent增加的特殊能力包括:1)对生成的代码进行反思和修改的能力;2)根据程序运行出错信息进行代码修改的能力。

1.3 大语言模型代理用于多问题生成

用户查询可能涉及多个交叉的研究方向,或者表述可能模糊不清。为了解决这一问题,我们采用了一个名为QuestionGenerationAgent的问题生成LLM智能体,根据用户输入生成5~10个高度相关的问题。这些问题随后被发送到向量数据库中进行检索。这种方法增强了搜索结果的多样性,能够从不

同学科中捕捉知识,并生成具有不同余弦相似度得分的检索结果。然而,这也可能检索到许多不相关的文章,因此需要 LLM 代理进一步进行语义分析。

QuestionGenerationAgent 具有进化的能力。最初生成的问题可能在相关性上存在差异。然而,通过多次迭代,当反思智能体 ReflectionAgent 提供关于检索文章相关性的反馈时,问题生成智能体会调整其策略以生成更好的搜索查询。在实际体验中,我们发现如果将智能体生成的问题列表中的问题一个个检索,则有些问题匹配的检索与最初的问题相关性较弱。此时,反思智能体的相关性反馈对问题生成智能体的提升起到较大帮助作用。但是,如果将生成的问题列表拼接成一个长语句进行检索,可以获得非常相关的检索信息,此时基本不需要反思智能体的介入。

实验发现,如果使用本地部署的 140 亿参数 DeepSeek-r1 来生成检索问题,效果不如使用 DeepSeek-r1 的官方 API 检索内容。下面是一个具体的例子,用户输入:

I want to study the nuclear equation of state, nuclear phase diagram, nuclear phase transition as well as critical end point in QCD phase diagram using high energy heavy ion collisions and deep learning, related studies can include machine learning approaches

DeepSeek-r1 经过推理分析,提出如下几个辅助问题:

"How does deep learning analyze high-energy heavy-ion collision data for nuclear equation of state?"

"What are recent advances in using machine learning to study nuclear matter properties from heavy-ion collisions?"

"Which deep learning techniques are applied to model the dense matter created in heavy ion collisions?"

"What role do observables like flow and fluctuation play in constraining the equation of state with machine learning?"

"How can neural networks be trained to extract transport properties from particle correlations in heavy-ion experiments?"

由于 arXiv 数据集是英文,用户的问题可能是中文,直接计算中文问题与英文摘要之间的向量相似度效果不好。用户直接使用英文提问,或者由问题生成智能体将用户的多语言转化为英文检索问题,与 arXiv 中的英文文章摘要拥有更好的相似度。

1.4 大语言模型智能体用于相关性分析

LLM Agent 生成的第一批问题可能在领域、研究问题或方法上与文章向量不完全匹配。相关性打分智能体对检索到的文章进行相关性分析、评分和排序。这一分析对于用户快速识别最相关和最重要的文献至关重要。此外,相关性分析还可以指导问题生成智能体生成更有效的搜索查询。

相关性打分智能体的评分可能并不总是与用户偏好一致。因此,用户可以手动对特定搜索结果进行评分,这些评分用于优化相关性打分智能体的判断,使其更符合用户意图。

下面是一个相关性打分智能体对检索出的文章打分的例子。之前的深度学习与核物质状态方程检索,向量数据库找到的第一篇文章^[2]题目如下:

Title: Exploring QCD matter in extreme conditions with Machine Learning

而相关性打分智能体经过判断,给出如下相关性打分以及解释:

相关性得分:1

这篇论文与用户的问题高度相关。两者均涉及利用机器学习技术研究极端条件下的 QCD 物质性质,并结合了重离子碰撞实验方法来探讨核的状态方程和相变过程等主题。

另外一篇非常相关的文章^[49]:

Title: Nuclear liquid-gas phase transition with machine learning

相关性分析:论文和用户问题的相关性得分是 0.95。原因是:

相同主题:两者都在研究核相变和使用机器学习方法结合实验数据。

相似方法:都涉及无监督和有监督的组合 learning 以及利用高能重离子碰撞的数据。

互补目标:用户关注的 QCD 物质与论文中的讨论直接相关。

1.5 结构化数据生成与修正智能体

在生成多个搜索查询时,我们希望 LLM 代理以 JSON 格式输出问题列表,例如 ["query1", "query2", ..., "queryn"]。然而,LLM 生成的 JSON 字符串常常由于 Markdown 标签、中文标点符号或其他格式错误而无法被 JSON 库解析。为了解决这一问题,我们引入了另一个代理 JSON 格式修正智能体,它首先使用正则表达式来纠正 JSON 字符串中的常见错误。在初步修正后,JSON 字符串及其解析错误会被传递给 JSON 格式修正智能体进行进一步优化。

为了彻底解决生成的 JSON 文本解析错误的情形,我们也尝试让问题生成智能体直接返回新的检索候选的字符串,对多个候选问题的拼接字符串进行向量化,从向量数据库中检索。这样的做法可以减轻 LLM 生成结构化数据时不能百分百解析的问题。

1.6 向量数据库与传统数据库的集成

虽然向量数据库在检索文章标题和摘要方面非常有效,但其他元数据(如 arXiv ID 和作者列表)可能更适合由传统数据库处理。因此,本系统将元数据(不包括摘要)存储在传统数据库 sqlite 中。当通过语义搜索检索到一篇文章时,其标题将用于查询传统数据库以获取更多信息。

传统数据库占据的硬盘空间远小于向量数据库,因为主要是文本信息,只占据 600 多 MB 磁盘空间。

1.7 智能 PDF 分析

一旦通过 arXiv ID 识别出高度相关的文章,即可下载其对应的 PDF 文件并上传至 DeepSeek-r1 等平台进行深入分析。或者,可以使用 Kimi 文献分析应用接口,通过网页链接 <https://papers.cool/arXiv/{id}> 获取详细见解。智能文献阅读接口通常回答关于给定文章的以下问题:

- Q1: 本文试图解决什么问题?
- Q2: 相关研究有哪些?
- Q3: 本文如何解决问题?
- Q4: 进行了哪些实验?
- Q5: 潜在的未来研究方向是什么?
- Q6: 总结本文的主要内容。

这些由 AI 生成的答案利用了 LLM 代理的检索增强生成(RAG)能力,提供了从 PDF 内容中提取的有用见解。比如 Kimi 的自动化文献阅读工具使用方法如下:已知文献的 arxiv id,可以使用网址 <https://papers.cool/arxiv/{arxiv id}> 打开一个网页。网页提供文献的标题、摘要、作者信息以及 4 个按钮,标签分别是 PDF、Copy、Kimi 和 REL。点击 Kimi 按钮,会调用大语言模型 Kimi 对相应的参考文献进行详细的解读。点击 REL 按钮,则可以列出与本研究文章相关的众多参考文献。每一篇参考文献都以同样的格式给出。

1.8 信息检索评估指标介绍

信息检索有多种评估指标,本文采用两种指标对基于向量数据库的语义检索能力进行评估:一种是查全率;一种是查准率。这两种指标的详细定义

和应用场景如下所述。

1) Recall@k(查全率)的定义

$$\text{Recall}@k = \frac{|\{\text{相关文档}\} \cap \{\text{前}k\text{个检索结果}\}|}{|\{\text{相关文档}\}|} \quad (2)$$

Recall@k 衡量系统能够在前 k 个结果中检索到的相关文档占有所有相关文档的比例。假设有 10 篇与查询相关的文章,系统在前 k 个结果中返回了其中的 5 篇,则 Recall@k 为 0.5 或 50%。这个指标告诉我们:系统能找回多少我们真正需要的文档。对于科研论文检索,查全率是非常重要的指标。

示例:查询"量子计算",存在 20 篇相关文章。如果系统在前 10 个结果中返回了 8 篇相关文章,则 Recall@10=8/20=0.4 或 40%。

2) Precision@k(查准率)的定义

$$\text{Precision}@k = \frac{|\{\text{相关文档}\} \cap \{\text{前}k\text{个检索结果}\}|}{k} \quad (3)$$

Precision@k 衡量前 k 个检索结果中有多少是相关文档。如果系统返回的前 10 个结果中有 7 个是相关的,那么 Precision@10 为 0.7 或 70%。这个指标反映了结果的纯净度:用户浏览结果时遇到相关内容的概率。

示例:系统返回的前 5 个结果中,有 4 个与查询相关,则 Precision@5=4/5=0.8 或 80%。

为了量化 arXiv 向量数据库的语义检索性能,我们构建了一个基准测试数据集。具体流程如下:首先,以"collision"为关键词,从每篇论文的标题和摘要(title+abstract)中进行检索,筛选出前 1 000 篇相关文献作为文档池。随后,我们设定检索查询语句:

Query="The effect of nuclear structure on heavy ion collisions."

通过调用 DeepSeek 云端 API,我们对这 1 000 篇文献进行相关性标注。标注标准为:若文献的标题和摘要内容与查询语句 Query 在主题和结论上具有显著相关性,则标注为"相关";否则标注为"不相关"。为确保标注准确性,除 DeepSeek 的自动评分外,我们还进行了人工复核验证。最终,同时被 DeepSeek 和人工标注为"相关"的文档将作为查全率和查准率计算中的{相关文档}集合。

在向量检索阶段,我们采用相同的嵌入模型将这 1 000 篇文献编码为向量表示,基于向量相似度进行语义检索,获取{前 k 个检索结果}。查全率和查准率的计算中,分子部分即为{前 k 个检索结果}与{相关文档}集合的交集。

2 AI 科学家系统使用方法介绍

该系统是一个基于 arxiv 向量数据库的 AI 科学家助理工具,主要功能包括:

- 1) 向量数据库: 将 arxiv 文献元信息转换为向量,并存储在向量数据库中。
- 2) 文献检索: 根据用户输入的问题生成检索问题,并从数据库中查询相关文献。
- 3) 相关性评分: 计算文献与用户问题之间的相关性得分。调用 Kimi api,寻找更多相关文献。
- 4) AI 文献阅读: 调用 Kimi api,对文献进行阅读,并生成摘要。
- 5) 文献摘要压缩: 对文献摘要进行压缩,提取关键信息。
- 6) 未来研究方向提取: 从文献中识别出可以进一步探索的研究点。

2.1 安装

- 1) 克隆项目
git clone https://gitee.com/lgpang/arxiv_vectordb.git
cd arxiv_vectordb
- 2) 创建虚拟环境(可选)
python -m venv ai_scientist_env
source ai_scientist_env/bin/activate # 在 Linux 或 MacOS 上
ai_scientist_env\Scripts\activate # 在 Windows 上
- 3) 安装依赖
pip install -r requirements.txt

2.2 使用

- 1) 启动应用
streamlit run streamlit.py
- 2) 在浏览器中访问: 打开浏览器并访问 http://localhost:8501。
- 3) 输入问题: 在应用界面中输入您的科研问题,点击“提交”按钮,系统将生成相关文献并显示结果。

2.3 依赖的数据文件

- 1) arxiv/arxiv-metadata-oai-snapshot.json: 包含 arXiv 文献的元数据,约 4.5 GB。
 - 2) docs/arxiv_articles.db: 传统的 SQLite 数据库,存储文献的相关信息,约 600 MB。
 - 3) docs/chroma/: 向量数据库,存储文献的向量信息,约 30 GB,请确保有足够的磁盘空间。
 - 4) bge-m3/: 向量嵌入模型,用于将文献元信息转换为向量,约 1.5 GB。
- 请确保这些数据文件和模型文件在项目的相应

目录中。

这些文件可以从百度网盘下载。

链接: <https://pan.baidu.com/s/1LufumbcuhX6XUbeIVdtrmQ?pwd=1121> 提取码: 1121。

2.4 获取 DeepSeek 的 API KEY

- 1) 访问 DeepSeek 官方网站。
 - 2) 注册一个账户并登录。
 - 3) 在用户设置中找到 API 密钥,复制并保存。
- 将 API 密钥添加到项目根目录下的 .env 文件中,格式如下:
- API_KEY=你的_api_key
BASE_URL=你的_base_url

2.5 部署本地大语言模型 DeepSeek-r1

- 1) 安装 Ollama: 请参考 Ollama 官方文档进行安装。
 - 2) 下载 DeepSeek 模型: 使用以下命令下载 DeepSeek-r1 模型。
ollama pull deepseek-r1:1.5b
 - 3) 启动 Ollama 服务: 启动 Ollama 服务以便本地访问。
ollama run deepseek-r1:1.5b
- 如果本地电脑配置比较好,比如有英伟达的显卡 RTX4090,可以使用 14b,即 140 亿参数的本地 DeepSeek 模型。需要注意的是,DeepSeek 的官方 API 在生成最初的检索问题候选时,效果远好于使用本地部署的小系统。因此本地部署的 DeepSeek-r1 只进行相关性分析。
- 4) 确保服务运行: 确保 Ollama 服务在 http://localhost:11434 上运行,以便系统能够访问。当然,也可以完全放弃本地部署 DeepSeek,直接使用 DeepSeek 的云端 api 进行相关性打分。

3 检索结果分析

为了量化比较语义检索与关键词检索,我们构建了一个信息检索评估基准数据库。这个数据库根据重离子碰撞领域的关键词 collision 找出 1 000 篇文章作为小的基准文献池,并使用 DeepSeek 辅助人工标注每篇文章与特定 Query("The effect of nuclear structure on heavy ion collisions.")的相关性,找出了 18 篇相关文献。接下来我们对比三种检索方法的查全率与查准率,这三种方法分别是:

- 1) 直接使用关键词[nuclear structure, heavy ion collision]在标题与摘要中检索。
- 2) 直接使用用户输入的 Query 在向量数据库中进行语义检索。

3)根据用户输入的Query,使用DeepSeek-r1产生一些检索候选Query,在向量数据库中检索。

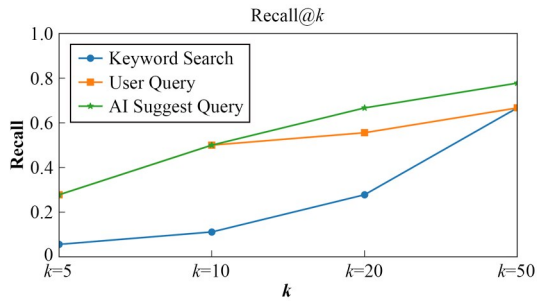


图2 三种不同检索方法的查全率(Recall@k)性能对比(彩图见网络版)

其中,蓝色圆点代表基于关键词的传统检索方法,橙色方块表示直接使用用户原始Query的语义检索结果,绿色五角星则呈现了通过DeepSeek生成Query扩展问题后基于语义检索的改进结果

Fig.2 The performance comparison of Recall@k among three different retrieval methods (color online)

The blue dots represent the traditional keyword-based retrieval method, the orange squares denote the semantic retrieval results using the user's original query directly, and the green pentagrams illustrate the improved semantic retrieval results achieved through query expansion questions generated by DeepSeek-r1.

图2展示了不同检索方法的查全率(Recall@k)性能对比。随着k值的增大,系统检索到的相关项增多(查全率普遍上升)。当k比较小时,基于关键词的查全率非常低,显著低于语义检索的查全率。当k=50时,在这个特定的数据集上,基于关键词检索的查全率与基于用户Query的查全率相等,都在65%左右。基于用户的Query,让DeepSeek生成更多更相关的查询语句,能够显著增加查全率。当k=50时,这种语义检索的查全率接近80%。因为总的相关文献为18篇,30%的查全率对应5篇,50%的查全率对应9篇,即k≤10的时候几乎所有基于语义检索获得的文章都是相关文章。因此基于语义检索对于大k时追求全面覆盖,小k时追求高质量结果拥有非常重要的意义。

图3展示了不同检索方法的查准率(precision@k)性能对比。随着k值的增大,系统检索到的相关文献增多,但返回结果中包含更多无关项,因此导致查准率下降。这是典型的查准率-查全率权衡关系。当检索的文献数量k<10时,直接使用用户Query和DeepSeek扩展的Query做语义检索基本无差别,都能达到100%(k=10)以及90%(k=1)的查准率。但是继续增加检索文献数量时,查准率会降低,比如k=20时,两种语义检索的查准率分别降低到60%和50%,当k=50时,查准率降低到20%。基于关键字的检索查准率一直在20%左右,意味着

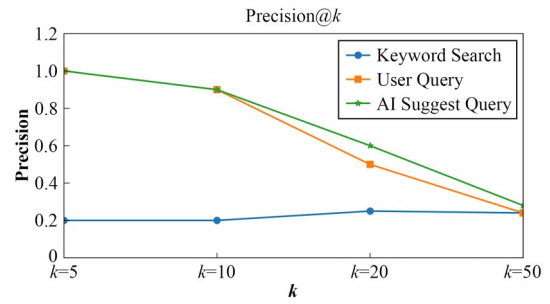


图3 三种不同检索方法的查准率(Precision@k)性能对比(彩图见网络版)

其中,蓝色圆点代表基于关键词的传统检索方法,橙色方块表示直接使用用户原始Query的语义检索结果,绿色五角星则呈现了通过DeepSeek生成Query扩展问题后基于语义检索的改进结果

Fig.3 The performance comparison of Precision@k among three different retrieval methods (color online)

The blue dots represent the traditional keyword-based retrieval method, the orange squares denote the semantic retrieval results using the user's original query directly, and the green pentagrams illustrate the improved semantic retrieval results achieved through query expansion questions generated by DeepSeek-r1.

k=5时只有1篇是相关文献,k=10时只有两篇相关文献。而基于语义检索的arxiv向量数据库,当k=5时,5篇检索出的文献都是相关文献;当k=10时,检索出的文献有9篇是相关文献。虽然看上去k=50的时候,语义检索的查准率也下降得比较厉害,但是这是一种错觉,因为查准率计算公式的分母是k,30%左右的查准率对应的相关文档数量为15篇,已经接近全部的相关文献数量(18篇)。

综合来看,AI推荐查询方法在所有k值下都明显优于关键词检索。关键词检索的查准率普遍较低,表明返回了较多不相关结果。关键词检索需要较高的k值才能获得可接受的查全率。如果可选,应优先使用AI推荐查询的语义检索方法;如必须使用关键词检索,建议使用较大的k值以保证足够的查全率,但需接受较低的查准率。对于向量数据库,可以根据具体应用需求及查准率/查全率的优先级,选择最合适的k值参数。

1)追求高质量结果:采用较小k值(如k=5或k=10)。

2)追求全面覆盖:采用较大k值(如k=20或k=50)。

3)平衡方案:k=10或k=20可在查准率与查全率间取得较优折中。

4 未来改进方案

4.1 未解决科学问题的识别

在对PDF文章进行深入分析时,可以使用未来

工作智能体来识别未解决的问题或未来研究方向。这是一项关键功能,因为一个科学家未解决的问题可能已被其他人解决或正在研究中。自动化检索这些问题可以弥合不同学科和研究人员的认知差距。这一步也可以由 Kimi 的自动文献阅读 API 提供,对应§3 中的 Q5:潜在的未来研究方向是什么?这些潜在的未来研究方向,可能已经被其他的研究者进行了充分的研究,或者能够从其他研究者的研究结果进行进一步的推理得到。使用 DeepSeek-r1 的强推理能力,从旧的知识库中提取关键信息,生成新的知识,是这个 AI 科学家系统的终极目标。这一步的实现可能需要更多的研究和创新性尝试。

4.2 基于时间序列的趋势分析

对于检索出来的相关文章,无论是通过硬编码的方式,还是通过人工智能的方式,都可以进行快速的时间排序,从而得到研究内容的时间序列。对于一个特定的研究领域,这样的时间序列还可以给出整个领域研究兴趣的自然转变,热点的出现与消亡,发现未解决但被遗忘的科学问题等等。这也是未来的 AI 科学家系统需要加入的功能。

4.3 文献综述的自动化生成

当向量数据库和大语言模型 Agent 帮助我们获取关于某个研究课题的最相关的参考文献,并自动化地判断文献的相关程度时,原则上可以根据时间序列、领域趋势分析、不同文献结论之间的异同、未解决问题的总结,完成特定领域的文献综述编写功能。这一步可能需要编写更多的大语言模型智能体,每个智能体负责完成不同的功能。在此过程中,DeepSeek 的云端 API 将起到非常重要的作用。作者在 2024 年 5 月开始使用 DeepSeek 的云端 API 完成一些科研辅助任务时,发现 DeepSeek 的云端 API 有并发功能,可以同时提交上千个问题,让 DeepSeek 同时并行地予以分析和回答。在未来,当 DeepSeek 的云端 API 负载能力可以再次承担并发任务时,我们希望这个 AI 科学家系统可以并发式地帮助用户分析领域的所有相关文献,并帮助撰写综述,或者基于已有的知识、自动发现的未解决科学问题,使用 DeepSeek-r1 的强推理能力,推理出新的知识,最终实现自动化的有价值知识生成。

4.4 多模态与动态知识库

目前的向量数据库主要使用 arxiv 元数据中标题与摘要信息,未来可以结合多模态分析,提取 PDF 源文件中更多的信息,结合文本、图表、代码(如 arXiv 的 LaTeX 源码)增强理解。开源的知识库只更

新到 2025 年 2 月底,为了跟踪最新的科研进展,需要构建动态知识库,实时更新 arXiv 数据,减少信息滞后。虽然 arxiv 向量数据库的搭建已经显著减少了 LLM 生成内容的幻觉风险,但未来我们需要引入更严格的事实核查机制,构建可受信任 AI 系统。

5 结语

本研究整合了 arXiv 平台上开源的 266 万余篇学术文献的标题与摘要数据,构建了规模达 30 GB 的向量数据库。基于向量余弦相似度计算并结合本地部署的 DeepSeek-r1 推理模型进行语义相关性分析,我们开发了一套高效的智能语义检索系统。通过对比关键词检索、用户 Query 检索以及 AI 建议检索三种方法的查全率和查准率,我们发现基于 arXiv 向量数据库的语义检索在所有 k 值下都明显优于关键词检索。该系统不仅适用于核物理领域研究,更能有效促进跨学科研究的开展,有助于揭示不同学科间的潜在关联。

我们将已完成嵌入处理的向量数据库开源,并存储在云盘中,便于相关领域研究者下载并在本地环境中部署使用。本项目还包含多个大语言模型智能体的构建示例。通过开源这些资源,我们旨在为领域研究者提供基础代码支持,助力开发更先进的科研辅助工具。这一举措不仅将加速核物理研究进程,还将促进跨学科知识的深度融合与创新发展。

致谢 感谢 AI 科学家系统的研究,让笔者可以自由开展人工智能在物理学中的应用。

作者贡献声明 本文相关的代码编写与文章撰写由庞龙刚独立完成。

参考文献

- 1 Akmal A, Pandharipande V R, Ravenhall D G. Arxiv dataset[EB/OL]. <https://www.kaggle.com/datasets/Cornell-University/arxiv>.
- 2 Zhou K, Wang L X, Pang L G, *et al.* Exploring QCD matter in extreme conditions with Machine Learning[J]. *Progress in Particle and Nuclear Physics*, 2024, **135**: 104084. DOI: 10.1016/j.pnpnp.2023.104084.
- 3 Mohanty B. Exploring the QCD phase diagram through high energy nuclear collisions: an overview[C]// *Proceedings of 8th International Workshop on Critical Point and Onset of Deconfinement — PoS(CPOD 2013. Napa, California, USA. Sissa Medialab, CPODCPOD2013: 001*. DOI: 10.22323/1.185.0001.
- 4 Mohanty B. Exploring the QCD phase diagram through

- relativistic heavy ion collisions[J]. EPJ Web of Conferences, 2014, **66**: 04022. DOI: 10.1051/epjconf/20146604022.
- 5 Du L P, Sorensen A, Stephanov M. The QCD phase diagram and Beam Energy Scan physics: a theory overview[J]. International Journal of Modern Physics E, 2024: 2430008. DOI: 10.1142/s021830132430008x.
- 6 Gelis F. Some aspects of the theory of heavy ion collisions [J]. Reports on Progress in Physics, 2021, **84**(5): 056301. DOI: 10.1088/1361-6633/abec2e.
- 7 Shen C. Dynamic modeling for heavy-ion collisions[J]. EPJ Web of Conferences, 2022, **259**: 02001. DOI: 10.1051/epjconf/202225902001.
- 8 Ohnishi A. Phase diagram and heavy-ion collisions: overview[J]. Progress of Theoretical Physics Supplement, 2012, **193**: 1 – 10. DOI: 10.1143/ptps.193.1.
- 9 Li X B, Guo R R, Zhou Y, *et al.* Machine learning phase transitions of the three-dimensional Ising universality class[J]. Chinese Physics C, 2023, **47**(3): 034101. DOI: 10.1088/1674-1137/aca5f5.
- 10 Meyer-Ortmanns H. Phase transitions in quantum chromodynamics[J]. Reviews of Modern Physics, 1996, **68**(2): 473 – 598. DOI: 10.1103/revmodphys.68.473.
- 11 Du Y L, Pablos D, Tywoniuk K. Jet tomography in heavy-ion collisions with deep learning[J]. Physical Review Letters, 2022, **128**: 012301. DOI: 10.1103/physrevlett.128.012301.
- 12 Du Y L, Zhou K, Steinheimer J, *et al.* Identifying the nature of the QCD transition in heavy-ion collisions with deep learning[J]. Nuclear Physics A, 2021, **1005**: 121891. DOI: 10.1016/j.nuclphysa.2020.121891.
- 13 Tomasik B. Ultrarelativistic heavy-ion collisions: exploring the phase diagram of QCD[J]. arXiv e-Prints, 2006: nucl-th/0610042. DOI: 10.48550/arXiv. nucl-th/0610042.
- 14 Tejeda-Yeomans M E. Heavy-ion physics: freedom to do hot, dense, exciting QCD[EB/OL]. 2020: 2004.13812. <https://arxiv.org/abs/2004.13812v1>.
- 15 Wang R, Ma Y G, Wada R, *et al.* Nuclear liquid-gas phase transition with machine learning[J]. Physical Review Research, 2020, **2**(4): 043202. DOI: 10.1103/physrevresearch.2.043202.
- 16 Gupta S. A short introduction to heavy-ion physics[J/OL]. arXiv, 2015, 1508.01136v2.
- 17 Du Y L, Zhou K, Steinheimer J, *et al.* Identifying the nature of the QCD transition in relativistic collision of heavy nuclei with deep learning[J]. The European Physical Journal C, 2020, **80**(6): 516. DOI: 10.1140/epjc/s10052-020-8030-7.
- 18 Kvasiuk Y, Zabrodin E, Bravina L, *et al.* Classification of equation of state in relativistic heavy-ion collisions using deep learning[J]. Journal of High Energy Physics, 2020, **2020**(7): 133. DOI: 10.1007/JHEP07(2020)133.
- 19 Nahrgang M. Current status of dynamical modeling of fluctuations at the QCD phase transition in heavy-ion collisions[J/OL]. arXiv, 2015, 1510.08146v1.
- 20 Guest D, Cranmer K, Whiteson D. Deep learning and its application to LHC physics[J]. Annual Review of Nuclear and Particle Science, 2018, **68**: 161 – 181. DOI: 10.1146/annurev-nucl-101917-021019.
- 21 Luo X F, Shi S S, Xu N, *et al.* A study of the properties of the QCD phase diagram in high-energy nuclear collisions [J]. Particles, 2020, **3**(2): 278 – 307. DOI: 10.3390/particles3020022.
- 22 Mohanty B, Xu N. Probe of the QCD phase diagram with ϕ -mesons in high-energy nuclear collisions[J]. Journal of Physics G: Nuclear and Particle Physics, 2009, **36**(6): 064022. DOI: 10.1088/0954-3899/36/6/064022.
- 23 Hartnack C, Puri R K, Aichelin J, *et al.* Modelling the many-body dynamics of heavy ion collisions: Present status and future perspective[J]. The European Physical Journal A - Hadrons and Nuclei, 1998, **1**(2): 151 – 169. DOI: 10.1007/s100500050045.
- 24 Stephanov M. QCD critical point: Recent developments [J]. EPJ Web of Conferences, 2024, **314**: 00042. DOI: 10.1051/epjconf/202431400042.
- 25 d'Enterria D. High-energy heavy-ions physics: from RHIC to LHC[J]. Nuclear Physics A, 2007, **782**(1 – 4): 215 – 223. DOI: 10.1016/j.nuclphysa.2006.10.038.
- 26 Sangaline E, Pratt S. Toward a deeper understanding of how experiments constrain the underlying physics of heavy-ion collisions[J]. Physical Review C, 2016, **93**(2): 024908. DOI: 10.1103/physrevc.93.024908.
- 27 Fries R J. High energy nuclear collisions: Theory overview[J]. Pramana, 2010, **75**(2): 235 – 245. DOI: 10.1007/s12043-010-0112-x.
- 28 Nahrgang M. The QCD critical point and related observables[J]. Nuclear Physics A, 2016, **956**: 83 – 90. DOI: 10.1016/j.nuclphysa.2016.02.074.
- 29 Abdughani M, Ren J, Wu L, *et al.* Supervised deep learning in high energy phenomenology: a mini review[J]. Communications in Theoretical Physics, 2019, **71**(8): 050009-10

955. DOI: 10.1088/0253-6102/71/8/955.
- 30 Lappi T. Initial state in heavy ion collisions[J]. Nuclear and Particle Physics Proceedings, 2016, **276 – 278**: 29 – 34. DOI: 10.1016/j.nuclphysbps.2016.05.007.
- 31 Gavin S. Summary of heavy ion theory[J/OL]. 1994 Meeting of the American Physical Society, Division of Particles and Fields (DPF 94), 1994: 1353 – 1360. hep-ph/9409328v2.
- 32 Fernández Madrazo C, Heredia Cacha I, Lloret Iglesias L, *et al.* Application of a convolutional neural network for image classification to the analysis of collisions in high energy physics[J/OL]. arXiv e-Prints, 2017: arXiv: 1708.07034. DOI: 10.48550/arXiv.1708.07034.
- 33 Du Y L, Pablos D, Tywoniuk K. Deep learning jet modifications in heavy-ion collisions[J]. Journal of High Energy Physics, 2021, **2021**(3): 206. DOI: 10.1007/JHEP03(2021)206.
- 34 Murase K, Akamatsu Y, Asakawa M, *et al.* Dynamically integrated transport approach for high-energy nuclear collisions at high baryon density[J]. JPS Conference Proceedings, 2019, **26**: 024016. DOI: 10.7566/JPSCP.26.024016.
- 35 Luo X F. Exploring the QCD phase structure with beam energy scan in heavy-ion collisions[J]. Nuclear Physics A, 2016, **956**: 75 – 82. DOI: 10.1016/j.nuclphysa.2016.03.025.
- 36 Petersen H. Beam energy scan theory: Status and open questions[J]. Nuclear Physics A, 2017, **967**: 145 – 152. DOI: 10.1016/j.nuclphysa.2017.04.018.
- 37 Hirvonen H, Eskola K J, Niemi H. Deep learning for flow observables in ultrarelativistic heavy-ion collisions[J]. Physical Review C, 2023, **108**(3): 034905. DOI: 10.1103/physrevc.108.034905.
- 38 Brandenburg J D. Exploring QCD matter at high baryon density[J]. Journal of Physics: Conference Series, 2020, **1667**: 012004. DOI: 10.1088/1742-6596/1667/1/012004.
- 39 Feng Z Q. Nuclear dynamics and particle production near threshold energies in heavy-ion collisions[J]. Nuclear Science and Techniques, 2018, **29**(3): 40. DOI: 10.1007/s41365-018-0379-z.
- 40 Wu S J, Shen C, Song H C. Dynamically exploring the QCD matter at finite temperatures and densities: a short review[J]. Chinese Physics Letters, 2021, **38**(8): 081201. DOI: 10.1088/0256-307x/38/8/081201.
- 41 Ma Y G, Pang L G, Wang R, *et al.* Phase transition study meets machine learning[J]. Chinese Physics Letters, 2023, **40**(12): 122101. DOI: 10.1088/0256-307x/40/12/122101.
- 42 Iancu E. QCD in heavy ion collisions[J]. ESHEP, 2011: 197 – 266. 1205.0579v1.
- 43 Li Z M. Overview of intermittency analysis in heavy-ion collisions[J]. Modern Physics Letters A, 2022, **37**(13): 2230009. DOI: 10.1142/s0217732322300099.
- 44 Pandav A, Mallick D, Mohanty B. Search for the QCD critical point in high energy nuclear collisions[J]. Progress in Particle and Nuclear Physics, 2022, **125**: 103960. DOI: 10.1016/j.pnpnp.2022.103960.
- 45 Rothkopf A. Bayesian techniques and applications to QCD [J]. PoS Confinement, 2018: 026. 1903.02293v1.
- 46 Frankfurt L, Strikman M. QCD: looking forward[J]. Surveys in High Energy Physics, 1999, **14**: 9 – 27. DOI: 10.1080/01422419908228837.
- 47 Mohanty B, Xu N. QCD critical point and high baryon density matter[J]. Acta Physica Polonica B Proceedings Supplement, 2021, **14**(2): 231. DOI: 10.5506/aphyspolb.2021.14.231.
- 48 Asakawa M, Kitazawa M, Müller B. Issues with the search for critical point in QCD with relativistic heavy ion collisions[J]. Physical Review C, 2020, **101**(3): 034913. DOI: 10.1103/physrevc.101.034913.
- 49 He W B, Li Q F, Ma Y G, *et al.* Machine learning in nuclear physics at low and intermediate energies[J]. Science China Physics, Mechanics & Astronomy, 2023, **66**(8): 282001. DOI: 10.1007/s11433-023-2116-0.
- 50 Vitev I. Probing the phases of qcd in ultra-relativistic nuclear collisions[J]. International Journal of Modern Physics A, 2005, **20**(16): 3777 – 3782. DOI: 10.1142/s0217751x0502759x.
- 51 Lappi T. Theory overview of Heavy Ion collisions[J/OL]. arXiv, 2016, 1609.04917v1. DOI: 10.1007/s12043-010-0112-x.
- 52 Pang L G, Zhou K, Su N, *et al.* An equation-of-state-meter of quantum chromodynamics transition from deep learning[J]. Nature Communications, 2018, **9**(1): 210. DOI: 10.1038/s41467-017-02726-3.
- 53 Chen J L, Xiao S T, Zhang P T, *et al.* BGE M3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation[EB/OL]. 2024: 2402.03216. <https://arxiv.org/abs/2402.03216v4>.
- 54 Andrew N. <https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/?ref=dl-staging-website.ghost.io>.