基于大模型语义对齐和判别的核动力文档搜索

徐浩然[1] 曹杰铭[2] 张瑾昀[1] 颜 雄[1]
XU Haoran CAO Jieming ZHANG Jinyun YAN Xiong

摘 要 针对核动力领域知识文档数量激增带来的查阅需求，传统搜索技术存在明显局限——字符串匹配难以适配专业术语的多样化表达，深度表征方法则受限于模型通用性与领域语义理解能力。为此，文章提出了一种基于大语言模型的核动力知识文档搜索技术，利用其世界知识储备与少样本学习能力，构建语义对齐与相关性判别的优化框架。首先，通过大语言模型的语义对齐算法，对用户查询、用户画像及核动力知识文档进行语义关联建模，缩小语义差距以提升建模精度；其次，采用大模型相关性计算算法优化传统深度学习模型的初步搜索结果，并结合少样本学习进一步提高相关性判别准确性。实验结果表明，该技术可显著提升核动力知识文档的检索效率与准确性，为核动力领域相关工作提供有效支撑。

关键词 大语言模型；文档搜索；核动力；语义建模

doi：10.3969/j.issn.1672-9528.2025.09.007

0 引言

随着国家对核动力基础设施的持续投资，核动力设施的设备数量和种类不断增加。这一趋势导致核动力相关单位积累了大量涵盖设计、研发、建造等全周期流程的知识文档，在一定程度上增加了工作人员查阅这些文档的时间成本，从而影响了核电建设工作的效率。但文档搜索技术作为一种广泛应用于不同知识查询场景的智能技术，可以有效缓解因知识文档过多而导致的查询低效问题[1]。因此，在核动力领域，开发高效且智能的核动力知识文档搜索技术显得尤为重要。目前，知识文档搜索技术旨在根据用户的查询输入和个人画像，从文档库中找出最相关的文档作为搜索结果。传统的搜索技术主要分为两类：基于字符串统计匹配方法[2]和基于深度学习表征[3]方法。字符串统计匹配方法通过寻找用户查询、用户画像与文档中完全相同的文本片段，并结合统计规律来判断文档的相关性。然而，在核动力领域，专业术语常常有多种表达方式。例如，“核电负荷”可能被称为“核电载荷”，使字符串匹配方法在核动力领域的效果有限。

近年来，基于深度学习表征的方法成为主流，其通常使用 BERT 等模型对用户查询、用户画像和知识文档进行向量化表示，然后通过计算用户向量和文档向量之间的语义关联性来判断文档相关性。然而，这种方法在核动力文档搜索中面临两大挑战：一是现有的深度学习模型通常只具备通用语言知识，而核动力文档中包含大量专业核领域术语，使模型难以准确建模用户查询及画像与核动力文档之间的语义关联性；二是深度学习模型需要大量标注数据进行监督训练，而核动力领域缺乏这一数据，限制了模型的性能。

随着大语言模型的快速发展，这些挑战得到了部分缓解。大语言模型丰富的世界知识使其能够准确地建模用户查询、画像和知识文档之间的语义关联性。此外，其强大的少样本学习能力使其能够在不依赖大量监督数据的情况下，仍实现精确的相关性建模[4]。

基于此，本文提出了一种基于大模型语义对齐和相关性判别的核动力知识文档搜索技术（LLM4Search）。

针对挑战一，该技术引入了基于大模型的语义对齐算法。该算法利用大语言模型对用户查询、用户画像和核动力知识文档进行语义对齐，缩小语义差距，从而提高基于深度学习表征的搜索技术建模语义关联的准确性。

针对挑战二，该技术采用基于大模型的相关性计算算法，利用少样本学习技术对深度学习表征方法的搜索结果进行二次相关性判别，以更精确地返回最相关的文档给用户。需要注意的是，由于大语言模型的推理效率较低，直接用其替代传统搜索技术不符合核动力文档搜索任务对实时运行效率的要求。

1. 中国核动力研究设计院 四川成都 610213
2. 四川大学 四川成都 610065
[基金项目] 面向复杂异构数据的会话式问答系统研究（62272330）

因此，本文选择将大语言模型与深度学习表征的搜索技术相结合，以在应对现有技术挑战的同时确保高效性。通过实验，本文验证了该方法在核动力知识文档搜索场景中的有效性，为核动力建设工作的高效推进提供了有力支持。此技术不仅在理论上拓展了大语言模型在专业领域应用的边界，也在实践中为核动力行业的知识管理和信息获取提供了新的解决方案。

1 文档搜索技术研究概况

当前的文档搜索技术主要划分为三大类：基于字符串统计匹配的方法、基于深度学习表征的方法，以及基于大语言模型的方法。

（1）基于字符串统计匹配的方法：依赖于在用户查询、用户画像与文档中寻找相同的文本片段，并结合 TF-IDF 等算法来量化文档与用户信息之间的相关性。然而，在核动力领域，专业术语常常存在多种表达方式，例如“核电负荷”可能被称为“核电载荷”，这限制了字符串匹配方法的有效性。

（2）基于深度学习表征的方法：通常利用 BERT 等模型对用户查询、用户画像和知识文档进行向量化表示，然后通过计算余弦相似度或欧氏距离来评估用户向量与文档向量之间的相似度，以此判断文档的相关性。然而，这些模型往往仅具备通用语言知识，而核动力文档中包含大量专业术语，导致难以准确建模用户查询及画像与核动力文档之间的语义关联。此外，深度学习模型通常需要大量标注数据进行有监督训练，而核动力领域缺乏这样的数据，限制了模型性能的提升。

（3）基于大语言模型的方法：为上述挑战带来了有效的解决方案。大语言模型拥有丰富的世界知识，使其能够更准确地建模用户查询、画像和知识文档之间的语义关联性。此外，其强大的少样本学习能力使其在不依赖大量监督数据的情况下，仍然能够实现较为精确的相关性建模。然而，大语言模型在核动力文档搜索领域的应用仍然较为稀缺。

2 基于大模型语义对齐和相关性判别的核动力知识文档搜索技术（LLM4Search）

本文提出了一种基于大语言模型的核动力知识文档搜索技术（LLM4Search），该技术通过语义对齐和相关性判别来提升文档检索的准确性。其核心目标是利用大模型缩小用户查询、用户画像与知识问答之间的语义差距，并在有限的训练数据下利用大语言模型提高文档相关性的判别精度。如图 1 所示，LLM4Search 由两个主要组件构成：基于大模型的语义对齐算法和基于大模型的相关性计算算法。

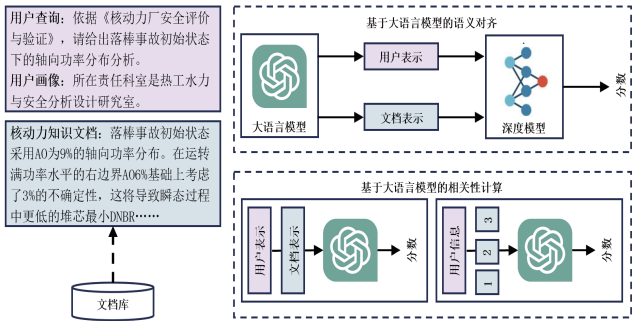

图 1 基于大模型语义对齐和相关性判别的核动力知识文档搜索技术（LLM4Search）

（注：该技术分为基于大模型的语义对齐算法和基于大模型的相关性计算算法）

由图 1 所示，首先，语义对齐算法通过大语言模型对用户查询、用户画像与核动力知识文档进行处理，以缩小语义差距。这一过程提升了后续深度学习表征方法在捕捉文档语义关联性方面的能力。随后，利用深度表征搜索技术进行初步检索后，相关性计算算法通过少样本学习技术对初始搜索结果进行进一步分析。即使在训练数据有限的情况下，该算法仍能有效识别与用户需求最相关的文档。考虑到 Qwen-72B[5] 的强大性能，本文在方法模块涉及大语言模型的部分都选择了 Qwen-72B 用于实现。

2.1 核动力知识文档搜索任务定义

对于用户 u_i，本文将用户的查询语句定义为 q_i，将用户的个人画像定义为 p_i，将深度模型定义为 $e(\cdot)$，将核动力知识文档库表示为 $H=\{h_1, h_2,\cdots, h_n\}$，其中 h_j 表示文档库中的第 j 个文档。

文档搜索任务旨在基于 q_i 和 p_i，计算文档库 H 中各文档 h_j 的相关性，并将相关性最高的 k 个文档返回给用户。

为了实现这一目标，基于深度学习表征的搜索方法会利用深度模型对用户查询、个人画像和知识文档进行向量化表示，然后计算向量之间的相似度，以判断文档的相关性。这一过程可以表示为：

$$s_j = \cos(\boldsymbol{e}(q_i+p_i), \boldsymbol{e}(h_j)) \quad (1)$$

式中：s_j 代表文档 h_j 的相关性分数；$\cos(\cdot)$ 代表向量相似度计算；$\boldsymbol{e}(q_i+p_i)$ 代表使用深度模型对用户查询 q_i 和画像 p_i 进行向量化表征；$\boldsymbol{e}(h_j)$ 代表使用深度模型对文档 h_j 进行表征。

计算结束后，相关性分数最高的 k 个文档将作为搜索结果返回给用户。

2.2 输入数据示例

本文提供了核动力知识文档搜索任务的输入数据示例，以更详细地展示该任务的输入输出设置。具体的数据示例如表 1 所示。

表 1 输入数据示例

字段名称	数据示例
用户画像 p_i	责任科室：热工水力与安全分析设计研究室 所属项目：福建福清核电厂 5、6 号机组 ……
用户查询 q_i	依据《核动力厂安全评价与验证》（HAD102/17），请给出落棒事故初始状态下的轴向功率分布假设，并说明保守性。
核动力知识文档 h_j	落棒事故初始状态采用 AO 为 9% 的轴向功率分布。在运行图满功率水平的右边界 AO 为 6%……这将导致……最小 DNBR，因此是保守的……

2.3 基于大模型的语义对齐算法

本章提出了一种基于大模型的语义对齐算法。该算法利用大语言模型对用户查询 q_i、用户画像 p_i 和核动力知识文档 h_j 进行语义对齐，缩小语义差距。这有助于在后续提升模型 $e(\cdot)$ 在式（1）中建模用户和文档之间语义关联性的能力。因此，本文深入探讨了两种利用大模型实现语义对齐的方法：大模型显式语义对齐、大模型隐式语义对齐。

显式语义对齐：相关研究表明，大语言模型凭借其强大的世界知识和推理能力，具备出色的语义概括能力。大模型能够利用自身的知识对给定文本进行推理、概括和重述，生成新的文本版本。由于预训练参数的影响，重述后的文本在语义空间上趋近于大模型的知识空间。基于这一特性，该算法尝试利用大语言模型的语义概括能力，对用户查询 q_i、用户画像 p_i 和核动力知识文档 h_j 三类信息进行重述，使这三类信息在语义空间上更接近大模型的预训练语义知识，从而实现语义对齐。本文通过多次验证，选择了 Prompt 构建形式用于语义对齐。该构建形式经验证具有最佳的语义对齐效果如表 2 所示。

表 2 大模型显式语义对齐 Prompt 构建

1. 针对用户查询 q_i 你是一位核动力领域的专家。现在有一位核动力工作人员，他想要了解的内容是：{q_i}。 请基于你的专业知识，理解并概括用户此时的意图。
2. 针对用户画像 p_i 你是一位核动力领域的专家。现在有一位核动力工作人员，他的背景信息是：{p_i}。 请基于你的专业知识，概括这个用户可能对哪些方面的核动力知识感兴趣。
3. 针对核动力文档 h_j 你是一位核动力领域的专家。现在有一篇核动力领域的知识文档：{h_j}。 请基于你的专业知识，概括文档的关键信息。

通过上述的 Prompt 构建，该算法可以对 q_i、p_i 和 h_j 分别进行语义重述。本文将重述后的信息定义为：q_i^*、p_i^* 和 h_j^*。随后，该算法将对齐后的信息通过式（1）输入给深度模型 $e(\cdot)$ 执行文档相关性的计算以实现搜索过程。

隐式语义对齐：隐式语义对齐不需要大模型显式地对输入文本进行概括和重述。相反，这种方法更像是将大语言模型用作编码器，替代传统的深度学习模型 $e(\cdot)$ 来进行文本编码，从而生成向量表示。随后，该算法可以直接计算大模型生成的文档向量与用户向量之间的相似度，以此计算文档相似度。这一过程为：

$$s_j = \cos(\mathrm{LLM}(q_i+p_i), \mathrm{LLM}(h_j)) \quad (2)$$

其中，在使用大模型对文本进行编码时，首先将待编码的文本输入到大模型中。然后，从大模型的最后一层网络中提取输出向量序列并进行平均处理，以生成文本的整体向量。

2.4 基于大模型的相关性计算算法

本算法充分利用了大模型的强大少样本学习能力，通过对深度搜索模型的相关性计算结果进行二次评估，以更精确地为用户返回最相关的文档。具体而言，本文采用语义对齐算法，结合式（1）和（2），首先从文档库中检索出 t 个初步相关的文档作为候选集，定义为：$H_i^t=\{h_1^i, h_2^i, \cdots, h_t^i\}$。然后，利用大模型结合用户信息，对候选集中的文档进行二次相关性计算，最终选出 k 个相关性最高的文档（其中 $k<t$）作为最终的搜索结果。此外。该算法使用了两种大模型相关性计算的 Prompt 构建范式：Point-Wise 判别式计算和 List-Wise 排序式计算。而在具体使用时，该算法会同时调用以上两种范式计算各文档的相关性并得到两种排序结果，并叠加得到各文档最终排序。

（1）Point-Wise 判别式计算：该算法直接将用户查询 q_i、用户画像 p_i，以及待计算的核动力知识文档 h_t^i 输入给大模型，要求大模型输出一个分数值来代表该文档的内容是否与用户的搜索意图相关。这一过程的 Prompt 如表 3 所示。

表 3 Point-Wise 判别式计算 Prompt 构建

请基于用户查询和用户画像理解用户的核动力知识搜索意图，并基于你的理解，输出一个 0~1 之间的分数来表示该核动力知识文档是否符合用户的搜索意图。 少样本示例：{...} 此时的输入： 用户查询及画像：q_i、p_i；文档内容：h_t^i 分数：

（2）List-Wise 排序式计算：该算法将用户查询 q_i、用户画像 p_i，以及整个待计算的核动力文档候选集 H_i^t 输入给大模型，要求大模型对 H_i^t 中各文档与用户信息的相关性程度进行排序，并输出一个排序后的顺序列表作为搜索结果。这一过程如表 4 所示。

表 4 List-Wise 排序式计算 Prompt 构建

请基于用户查询和用户画像理解用户的核动力知识搜索意图，对候选集中的文档按照与用户意图的匹配程度进行排序，并输出序号。 少样本示例：{...} 此时的输入： 用户查询及画像：q_i、p_i； 文档候选集：[1]h_1^i[2]h_2^i... [t] h_t^i 排序结果：

3 实验设置

3.1 数据集

本研究构建了核动力知识领域专用的中文核动力知识搜索数据集（CN-NUC），数据集收集了某核电厂系统从 2021 年 3 月—2024 年 3 月所查询的反应堆一回路系统中主要设备设计文本。这些文本包含了脱敏的核动力工作人员画像、查询语句以及核动力文档。该数据集总共包含 2 000 条数据，其中 1 800 条数据用于模型微调，200 条数据用于测试。

3.2 基线方法

本文选择了基于字符串统计匹配的方法 BM25，基于深度学习表征的方法 BERT、RoBERTa[6] 和 monoT5[7]，以及基于大语言模型的方法 Qwen 结合少样本学习 [8] 和检索增强 [9]，得到 QwenFewShot 和 QwenRAG。具体地，对于 BERT、RoBERTa 和 Qwen，本文采取了微调和不微调两种范式。此外，由于核动力文档通常具有较高的保密级别，本文未选择 ChatGPT、GPT4[10] 等未开源的大语言模型作为基线。

3.3 评估指标

本文选择了 3 个被广泛应用于评估搜索模型性能的评估指标：nDCG@1、nDCG@5 和 nDCG@10[11]。这些指标可以有效衡量搜索模型搜索正确结果的排序情况，值越高，表示模型搜索性能越好。

3.4 总体实验结果

从表 5 中可以看出，本文提出的方法 LLM4Search 在所有数据集上均表现出色，取得了最佳效果。特别是与表现最好的基线方法相比，利用隐式语义对齐的 LLM4Search 在 3 个指标上分别提升了 +2.49、+10.07 和 +7.29。而利用显式语义对齐的 LLM4Search 则提升了 +2.68、+10.36 和 +8.69，这表明本文的方法在核动力搜索任务中具有显著的性能优势。同时，使用显式语义对齐的 LLM4Seach 相较于隐式语义对齐，在 3 个指标上都有一定的优势。这说明了通过要求 LLM 显式地对文本内容进行概括和输出，可以更好地拉近用户信息和文档内容的语义关联。

表 5 总体实验结果

模型 \ 指标	nDCG@1	nDCG@5	nDCG@10
不微调（Non-Fine-Tuned）			
BM25	43.33	45.96	45.77
BERT	47.11	47.92	48.23
RoBERTa	48.52	49.07	50.26
QwenFewShot	54.29	52.93	55.02
QwenRAG	55.39	54.30	55.28
微调（Fine-Tuned）			
BERTFinetuned	56.38	53.05	56.44
RoBERTaFinetuned	57.64	54.83	56.52
monoT5	63.33	57.46	61.27
LLM4Search w/ 隐式语义对齐	65.82	67.53	68.56
LLM4Search w/ 显式语义对齐	**66.01**	**67.82**	**69.96**

3.5 消融实验

本文通过消融实验验证了所提出的大模型语义对齐和大模型相关性计算两个模块的有效性。实验结果如表 6 所示，去除大模型的语义对齐和相关性计算功能后，LLM4Search 的性能显著下降。具体来说，当消融语义对齐算法时，实验不再要求大模型对用户和文档信息进行概括性对齐，而是直接使用未处理的原始文档进行相关性计算。这一改变导致 LLM4Search 的性能显著下降，从而进一步证明了预先使用大模型进行语义对齐的重要性。在消融大模型相关性计算时，实验分别对两种相关性计算范式进行了测试。结果表明，每种计算范式的消融都导致了性能的一定程度下降，这表明这两种范式均对 LLM4Search 的性能提升有积极影响。

表 6 消融实验结果

模型 / 指标	nDCG@1	nDCG@5	nDCG@10
LLM4Search	**66.01**	**67.82**	**69.96**
w/o 语义对齐	64.53	63.19	67.62
w/o 大模型相关性计算 -PointWise	65.49	64.82	68.91
w/o 大模型相关性计算 -ListWise	64.76	63.35	67.74

4 结论

本文提出了一种基于大语言模型进行语义对齐和相关性判别的核动力知识文档搜索技术，有效解决了核动力领域知识文档搜索中的两大主要挑战。首先，通过采用大语言模型进行语义对齐，本文成功缩小了用户查询、用户画像与核动力知识文档之间的语义差距。这种方法提高了深度学习表征技术在建模语义关联性方面的准确性。其次，利用大语言模型的少样本学习能力，本文对搜索结果进行了二次相关性判别，即便在缺乏大量标注数据的情况下，依然能够实现精确

的文档相关性判断。实验结果表明，LLM4Search 在核动力知识文档搜索场景中表现出色，不仅提高了文档查找的效率，还加速了核电建设工作的整体推进。这一技术的成功应用，不仅扩展了大语言模型在专业领域应用的理论边界，也为核动力行业的知识管理和信息获取提供了创新的解决方案。

参考文献：

[1] 李金忠，刘伟东，陈盛博 . 搜索结果多样化排序 : 新进展与展望 [J/OL]. 计算机工程与科学 ,1-26[2025-01-17].http://kns.cnki.net/kcms/detail/43.1258.tp.20241011.1311.006.html.

[2]AIZAWA A. An information-theoretic perspective of TF–IDF measures[J]. Information processing & management, 2003, 39(1): 45-65.

[3]ZHAN J T, MAO J X, LIU Y Q, et al. Optimizing dense retrieval model training with hard negatives[EB/OL]. (2014-04-16)[2025-06-21].https://doi.org/10.48550/arXiv.2104.08051.

[4] 刘华玲，张子龙，彭宏帅 . 面向闭源大语言模型的增强研究综述 [J]. 计算机科学与探索 , 2025,19(5):1141-1156.

[5]YANG A, YANG B S, HUI B Y, et al. Qwen2 technical report[EB/OL].(2024-09-10)[2025-03-23].https://doi.org/10.48550/arXiv.

[6] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL].(2019-07-26)[2025-06-23].10.48550/arXiv.1907.11692.

[7]NOGUEIRA R, JIANG Z Y, LIN J. Document ranking with a pretrained sequence-to-sequence model[EB/OL]. (2023-03-14)[2025-06-12].https://doi.org/10.48550/arXiv.

[8]BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. NewYork: ACM, 2020: 1877-1901.

[9] 王合庆，魏杰，景红雨，等 .Meta-RAG: 基于元数据驱动的电力领域检索增强生成框架 [J/OL]. 计算机工程 ,1-11[2025-01-17].https://doi.org/10.19678/j.issn.1000-3428.0070415.

[10]ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB/OL].(2023-03-15)[2025-04-12].https://www.semanticscholar.org/paper/GPT-4-Technical-Report-Achiam-Adler/163b4d6a79a5b19af88b8585456363340d9efd04.

[11]JARVELIN K, KEKÄLÄINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM transactions on information systems (TOIS), 2002, 20(4): 422-446.

【作者简介】

徐浩然（1990—），男，四川南充人，硕士，馆员，研究方向：计算机在核动力方面的研究及应用。

曹杰铭（2000—），男，四川广安人，硕士研究生，研究方向：大语言模型，email:2416729460@qq.com。

张瑾昀（1996—），男，四川遂宁人，硕士，研究方向：数字化研发、知识图谱。

颜雄（1992—），男，湖北监利人，硕士，研究方向：计算机应用。

（收稿日期：2025-04-14 修回日期：2025-09-09）

（上接第 26 页）

参考文献：

[1] 潘今一，王亚蒙，王伟，等 . 基于风格迁移和薄板样条的扩充汉字样本方法 [J]. 浙江工业大学学报 ,2020, 48(1):25-29.

[2] 秦嘉霖，刘维尚 . 基于直观汉字构形原理的 C^3-GAN 字体生成优化方法 [J]. 包装工程 , 2023, 44(10):193-201.

[3] 姚伟健，赵征鹏，普园媛，等 . 稠密自适应生成对抗网络的爨体字风格迁移模型 [J]. 计算机辅助设计与图形学学报 , 2023, 35(6):915-924.

[4] 陈二开，李成城，邬友，等 . 基于改进 CycleGAN 的粉笔字书写风格迁移研究 [J]. 印刷与数字媒体技术研究 , 2024(6):100-109.

[5] 张攀，周芳利，杨彪，等 . 大千字库智能化构建方法研究 [J]. 内江师范学院学报 ,2024,39(6):55-59.

[6]zi2zi[EB/OL].(2019-06-26)[2025-02-16].https://github.com/kaonashi-tyc/zi2zi.

[7]ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2017: 1125-1134.

[8]ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV).Piscataway:IEEE,2017:2242-2251.

[9]CHANG B, ZHANG Q, PAN S Y, et al. Generating handwritten chinese characters using CycleGAN[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway:IEEE,2018: 199-207.

【作者简介】

张攀（1989—），男，四川资阳人，硕士，讲师，研究方向：交叉学科、图像处理。

徐瑞娟（2004—），女，江西瑞昌人，本科在读，研究方向：产品设计。

（收稿日期：2025-05-08 修回日期：2025-09-16）