ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 603－615
DOI:10. 16337/j. 1004-9037. 2025. 03. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于大语言模型的航空发动机领域高质量数据集构建

邹冠沄[1,2]，王存俊[3]，孔寅豪[3]，马小庆[3]，李丕绩[1,2]

(1. 南京航空航天大学人工智能学院,南京 211106;2. 模式分析与机器智能工业和信息化部重点实验室(南京航空航天大学),南京 211106;3. 中国商用飞机有限责任公司上海飞机设计研究院,上海 201210)

摘　要: 随着人工智能技术的快速发展,大语言模型(Large language models, LLMs)在多个领域的应用日益广泛。然而,航空发动机领域由于缺乏高质量的人工编写问答数据集,限制了专家问答大模型的应用。本文提出了一种基于LLMs的问答数据集自动化构建方法,该方法无需人工干预即可生成高质量的开放式问答数据。在数据生成阶段,采用上下文学习方法和输入优先生成策略,增强了生成数据的稳定性;在数据过滤阶段,通过原文相似度的忠实度评估和大模型的语义质量评估,建立了数据质量自动评估机制,有效筛选出受幻觉影响的异常数据,确保数据的事实可靠性。实验结果表明,该方法显著提升了生成数据集的质量,经过指令微调后的模型在航空发动机领域的知识问答表现显著提升。本文的研究成果不仅为航空发动机领域的大模型应用提供了坚实基础,也为其他复杂工程领域的数据集自动化构建提供了参考。

关键词: 大语言模型;垂直领域大模型;问答数据生成;问答数据质量评估

中图分类号: TP391　　**文献标志码:** A

Construction of High-Quality Dataset in Aero-engine Domain Based on Large Language Model

ZOU Guanyun[1,2], WANG Cunjun[3], KONG Yinhao[3], MA Xiaoqing[3], LI Piji[1,2]

(1. College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China;2. MIIT Key Laboratory of Pattern Analysis and Machine Intelligence (Nanjing University of Aeronautics and Astronautics), Nanjing 211106, China;3. COMAC Shanghai Aircraft Design &. Research Institute, Shanghai 201210,China)

Abstract: With the rapid advancement of artificial intelligence technology, large language models (LLMs) are increasingly being applied across various domains. However, the lack of high-quality, manually curated question-answering datasets in the field of aero-engine has hindered the practical application of expert-level question-answering model. To address this issue, this paper proposes an automated method for constructing question-answering datasets based on LLMs, which generates high-quality open-domain question-answering data without human intervention. During the data generation phase, the method employs in-context learning and input-priority generation strategies to enhance the stability of the generated data. In the data filtering phase, a dual evaluation mechanism is established, combining faithfulness assessment based on source text similarity and semantic quality evaluation using large language models, to automatically filter out hallucinated or anomalous data and ensure factual reliability. Experimental results demonstrate that the proposed method significantly improves the quality of the generated dataset. Models fine-tuned on this dataset exhibit notable performance improvements in aero-engine domain knowledge

收稿日期:2024-10-13;**修订日期:**2025-01-15

question-answering tasks. The findings of this study not only provide a solid foundation for the application of large language model in the aero-engine domain but also offer valuable insights for automated dataset construction in other complex engineering fields.

Key words: large language model; vertical domain large language model; question-answering data generation; quality assessment of question-answering data

引　言

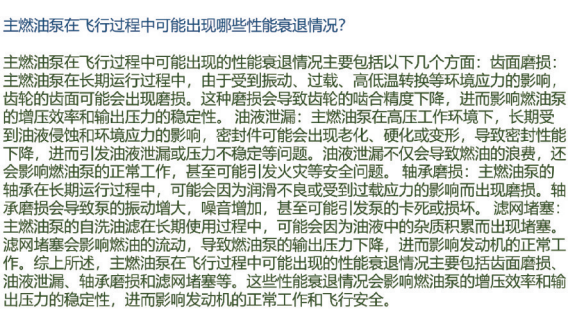
随着大语言模型(Large language model,LLM)领域近年来的高速发展,以ChatGPT系列[1]、LLaMA系列[2]和Qwen系列[3]等为代表的大模型,在知识问答、文本摘要和内容生成等自然语言处理任务上表现出卓越的通用性能[4]。这些模型不仅成为学术界的研究热点,也被工业界广泛应用于人工智能系统的构建。尽管通用大模型通过大规模预训练掌握了丰富的开放领域知识,但在处理垂直领域的复杂问题时,仍面临专业知识覆盖不足的挑战[5]。为此,文献[6]提出了提示工程,基于外部知识库的检索增强生成[7]及基于高质量数据集的监督微调[8]等方法,以提升模型在垂直领域的对话表现,使其能够更准确地理解用户输入并生成专业内容。然而,垂直领域大模型的发展面临高质量数据集匮乏的挑战。构建此类数据集通常依赖耗时且成本高昂的人工编写,而垂直领域文档的专业性进一步增加了对专业人员的需求,显著提升了人力成本和构建难度。特别是在航空发动机等工程领域,除公开的教材和论文外,还存在大量私有文档(如事故报告、设计文档和操作手册)。人工编写不仅难以全面覆盖领域知识,还可能因私有文档的保密性而受限。因此,利用大模型自动生成数据以替代人工编写,成为解决数据集短缺问题的有效途径。大模型在生成问答数据时,常因幻觉问题[9]及上下文遗忘[10]等缺陷,导致生成内容偏离事实或前后矛盾,严重影响数据集质量,因此亟需有效的数据质量度量方法对生成数据进行筛选。此外,封闭式问答(如多选题和判断题)难以满足人机对话场景的指令微调需求,而开放式问答数据的质量评估仍是一个未解决的难题。传统方法依赖人工评估筛选生成数据,但其效率与成本和人工编写数据集类似,存在显著局限性。

为解决上述问题,本文提出了一种基于大语言模型的高质量问答数据集构建方法,并构建了航空发动机领域问答数据集EngineQA。该方法通过上下文学习与分阶段生成策略优化提示工程,并结合基于原文相似度的忠实度评估与大模型驱动的语义质量评估,建立了开放式问答数据的自动评估机制,有效过滤低质量数据。本文的主要贡献总结为:

(1) 提出了一种基于大语言模型的高质量问答数据集生成方法。针对航空发动机领域数据集构建难题,通过上下文学习与输入优先生成策略优化生成过程,显著提升了数据的稳定性与一致性。

(2) 设计了开放式问答数据质量的自动评估机制。结合基于原文相似度的忠实度评估与大模型驱动的语义质量评估,有效过滤低质量数据,确保数据集的事实可靠性与语义准确性。

(3) 构建了航空发动机领域问答数据集EngineQA。该数据集包含8 800余条复杂工程问题及回答,覆盖领域专业知识,其开放式问答形式更贴近人机对话场景,适用于领域专家问答模型的指令微调,为航空发动机领域大模型应用提供了重要数据基础。EngineQA数据集的问答数据示例如图1所示。


主燃油泵在飞行过程中可能出现哪些性能衰退情况?

主燃油泵在飞行过程中可能出现的性能衰退情况主要包括以下几个方面：齿面磨损：主燃油泵在长期运行过程中，由于受到振动、过载、高低温转换等环境应力的影响，齿轮的齿面可能会出现磨损。这种磨损会导致齿轮的啮合精度下降，进而影响燃油泵的增压效率和输出压力的稳定性。 油液泄漏：主燃油泵在高压工作环境下，长期受到油液侵蚀和环境应力的影响，密封件可能会出现老化、硬化或变形，导致密封性能下降，进而引发油液泄漏或压力不稳定等问题。油液泄漏不仅会导致燃油的浪费，还会影响燃油泵的正常工作，甚至可能引发火灾等安全问题。 轴承磨损：主燃油泵的轴承在长期运行过程中，可能会因为润滑不良或受到过载应力的影响而出现磨损。轴承磨损会导致泵的振动增大，噪音增加，甚至可能引发泵的卡死或损坏。 滤网堵塞：主燃油泵的自洗油滤在长期使用过程中，可能会因为油液中的杂质积累而出现堵塞。滤网堵塞会影响燃油的流动，导致燃油泵的输出压力下降，进而影响发动机的正常工作。综上所述，主燃油泵在飞行过程中可能出现的性能衰退情况主要包括齿面磨损、油液泄漏、轴承磨损和滤网堵塞等。这些性能衰退情况会影响燃油泵的增压效率和输出压力的稳定性，进而影响发动机的正常工作和飞行安全。

图1　EngineQA数据集问答数据示例

Fig.1　Example of question-answering data from the EngineQA dataset

1 相关工作

(1) 问答数据生成。问答数据生成在专家问答与教育应用中具有重要意义,相关研究已取得一定进展。早期,Mitkov等[11]通过句法转换模板将文本转换为问答形式;Heilman等[12]结合手动排序规则与逻辑回归模型生成问答,提升了数据的人工评估认可度;Lee等[13]采用变分自编码器生成通用领域的高一致性问答数据。近期,Virani等[14]提出了多功能问答生成系统,能生成特定领域的多样化问答数据,但仍依赖预定义回答。本文利用大模型从特定领域文档中生成高质量的问答数据,无需人为预定义与评估筛选,生成的开放式闭卷问答数据形式更契合垂直领域专家问答系统的微调训练需求。

(2) 问答数据质量评估。评估长篇生成文本(如开放式问答对)的质量因主观性强而颇具挑战性。传统机器翻译指标如BLEU(Bilingual evaluation understudy);ROUGE(Recall-oriented understudy for gisting evaluation);METEOR(Metric for evaluation of translation with explicit ordering)依赖n-gram计算,难以捕捉回答与原文的语义相似度[15]。随着大模型的发展,其在语义评估中的潜力逐渐显现。Song等[16]通过实验验证了大模型在开放式问答数据评估中的有效性;Yue等[17]利用提示工程方法实现了问答数据归因性的自动评估。Wan等[18]设计了基于大模型的RACAR(Relevance, Agnosticism, Completeness, Accuracy, Reasonableness)指标,全面衡量科学问答数据的语义质量。本文结合数据生成策略与垂直领域对事实可靠性的高要求,设计了基于原文相似度的忠实度评估与大模型驱动的语义质量评估方法,从两个维度全面评估垂直领域开放式问答数据的质量。

(3) 垂直领域问答数据集。垂直领域问答基准数据集通常由科学文献、教科书或专家问答记录构成,对评估问答系统在垂直领域的性能至关重要。现有数据集多集中于医学与法律领域,且形式局限于选择题或判断题,而非开放式问答。例如,Jin等[19]构建的PubMedQA数据集基于PubMed文章摘要,测试问答系统通过生物医学文本推理回答问题的能力,但答案仅限于“是”“否”或“可能”三元分类。Guha等[20]构建的LegalBench数据集涵盖法律领域多种推理任务,旨在评估大模型在法律语料上的阅读理解与推理能力,但同样局限于二元分类。本文构建的EngineQA数据集采用开放式问答形式,更贴近人机交互场景,同时填补了航空发动机领域数据集的空白。本文数据集与其他垂直领域数据集的基本信息对比如表1所示。

表1 垂直领域数据集概览

Table 1 Overview of vertical domain datasets

数据集名称	构建方式	问答形式	数据量/10^3
PubMedQA	人工	三元判断	1
LegalBench	人工	二元判断	90
MedQA[21]	半自动	多项选择	10.2
EngineQA	自动	闭卷开放式	8.8

2 本文方法

2.1 问题定义及方法概览

大模型在文本摘要、抽取式问答等传统NLP任务中已展现出卓越性能,而生成问答对的任务可转化为依赖摘要与抽取能力的任务[22]。本文将问答数据生成任务形式化定义如下:对于包含领域知识的文档上下文片段c_i,大模型M需概括c_i的信息生成若干问题,即$M(c_i)=\{q_i\}_{i=1}^{n}$ ($n\leqslant 3$);随后,大模型M需理解c_i并从中抽取对应的事实信息,整合生成回答a_i,即$M(c_i, q_i)=a_i$。值得注意的是,在航空发动机领域,问题q_i不仅限于“是什么”或“什么时间”等简单询问,更多涉及“如何解决”或“原因是什么”等需要归纳推理能力的复杂问题,如图2所示。

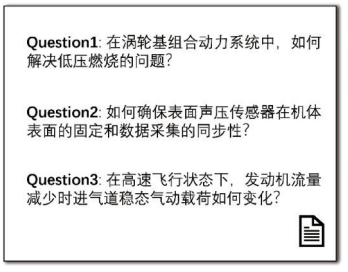


图2 生成复杂问题示例

Fig.2 Example of generating complex questions

图3为本文提出的一种基于大语言模型的数据集自动构建方法,包含3个阶段:语料预处理、数据生成和数据过滤。在语料预处理阶段,首先收集领域高质量语料,利用OCR(Optical character recogni-

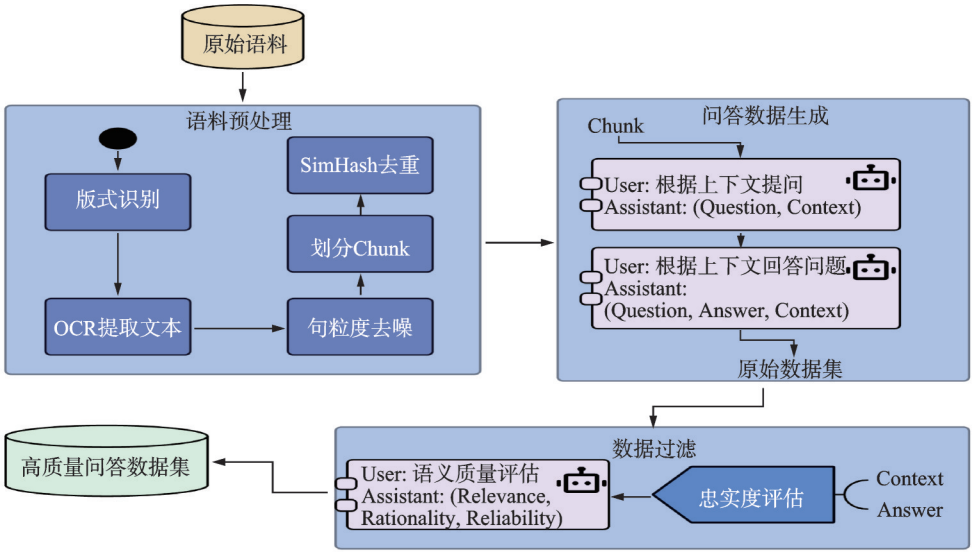


图 3 问答数据集自动构建流程

Fig.3 Flowchart of automatic construction process of question-answering dataset

tion)技术从PDF文档中提取文本,并进行去噪和去重等预处理操作。在数据生成阶段,采用提示工程中的上下文学习方法[23]与输入优先策略[24],分阶段生成问答数据。在数据过滤阶段,结合基于原文与回答嵌入向量余弦相似度的忠实度计算指标,以及基于大模型的语义质量评估指标,全面度量数据质量以过滤低质量数据。各阶段的实现细节算法表示如算法1所示。

算法1 问答数据集自动构建

输入:

D: 语料库,表示为文档集合 $D=\{d_1,d_2,\cdots,d_n\}$,其中 d_i 表示第 i 个文档

M: 生成大模型,用于生成问题和答案

E: 评估大模型,用于计算RAR(Relevance, Adequacy, Reliability)得分

F: 余弦相似度计算函数,用于计算回答与文档块之间的忠实度得分

T_r: RAR得分通过阈值

T_f: 忠实度得分通过阈值

输出:

H: 高质量问答数据集,表示为三元组集合 $H=\{(q_1,a_1,c_1),(q_2,a_2,c_2),\cdots,(q_n,a_n,c_n)\}$,其中 q_i 表示第 i 个问题,a_i 表示第 i 个回答,c_i 表示第 i 个文档块

(1) ➢语料预处理

(2) 对语料库 D 中的每个文档 d_i 进行句子粒度的文本去噪处理

(3) 去噪后的文档 d_i 划分为文档块,并存储在集合 $C=\{c_1,c_2,\cdots,c_L\}$ 中

(4) 对集合 C 进行SimHash去重处理,去除重复文档块

(5) ➢问答数据生成

(6) 初始化空集 H 用于存储生成问答数据

(7) for each $c_i \in C$ do

(8) 生成问题 q_i: $q_i \leftarrow M(c_i)$

(9) 生成回答 a_i: $a_i \leftarrow M(q_i, c_i)$
(10) 将三元组(q_i, a_i, c_i)插入集合$\boldsymbol{H}$
(11) end for

(12) ➢数据质量过滤
(13) for each $(q_i, a_i, c_i) \in \boldsymbol{H}$ do
(14) 计算RAR得分 r: $r \leftarrow E(q_i, a_i, c_i)$
(15) 计算忠实度得分 f: $f \leftarrow F(a_i, c_i)$
(16) if $r < T_r$ or $f < T_f$ then
(17) 将三元组(q_i, a_i, c_i)从集合$\boldsymbol{H}$删去
(18) end if
(19) end for
(20) return $\boldsymbol{H}$

2.2 语料预处理

本文收集的原始语料包括经典教材与中文期刊两部分。经典教材来源于西北工业大学姜长英数据库,涵盖《航空发动机原理》《航空发动机结构设计分析》等64本教材;中文期刊部分选自《航空发动机》(2001年至今,共128期)与《航空工程进展》(2019年至今,共27期)。所收集的语料覆盖航空发动机领域的基础理论、材料科学、控制系统及维护修理等多方面知识,确保了领域知识的完备性。

2.2.1 文本提取

由于初始语料均为PDF格式,本文使用PyMuPDF库将PDF切分为图片。鉴于期刊论文存在单列与双列两种排版方式,统一OCR处理可能导致段落不连贯,因此通过PyMuPDF计算文本块边界框宽度,准确识别排版方式并采用不同切分方法,确保单图片内语句的连续性。随后,使用PaddleOCR[25]开源OCR库从切分图片中提取纯文本信息。实验表明,与TesseractOCR相比,PaddleOCR在中文图片文本识别中具有更高的准确性。

2.2.2 文本去噪去重

通过OCR提取的非结构化文本常包含无意义的噪声文本,如参考引用、截断公式、书目号及页脚等数字和符号信息。为在去噪的同时尽量保留原文段落完整性,本文采用基于正则表达式的启发式规则过滤方法,以句子为单位处理不同类型噪声。具体而言,当中文字符比例远低于其他字符时,视为有效信息过少而丢弃该句;若句子涉及图片描述且后续生成输入不包含图片,则同样弃置;若句子仅含少量符号字符,则仅剔除符号而保留句子。尽管该方法尽可能保留了连贯段落信息,但仍不可避免地损失部分数值信息。

对去噪后的非结构化文本,需将其划分为固定长度的上下文文本块。为保证文本块内语义完整性,本文以句子为单位,按固定最小长度划分文本块,而非传统的固定长度与重叠长度划分方式。在划分长度选择上,本文通过实验评估了不同长度对提示工程效果及问答信息覆盖率的影响,最终确定以大于600字符(平均330 Tokens)作为划分长度。

为避免重复数据导致模型训练过拟合,本文采用基于SimHash[26]的相似度计算方法对文本块集合进行去重。SimHash通过提取文本特征,将高维特征向量映射为低维向量并转换为固定长度哈希值,从而获得长文本集合中每个元素的哈希表示。通过计算哈希值间的汉明距离,若距离低于经验阈值,则判定文本相似。由于文本块划分长度大于600字符,文本块数量达26 000余个,SimHash方法在处理长文本哈希去重时兼具高准确性与高效性,契合该阶段的大规模文本块去重需求。

2.3 问答数据生成

为避免生成问题复杂度过高，本文在提示词中强调问题的原子性与简洁性，确保大模型在后续回答中能准确抽取相关原文信息。同时为防止解答信息超出上下文范围，采用输入优先策略，要求模型基于原文实体生成问题。在本文实验过程中发现，在基于提示工程的问题生成场景中，多样本提示词的表现逊于零样本提示词。这是因为问题生成对输出稳定性要求较低，而更注重提问方式的多样性，以确保问答数据集在对话场景中的有效性，因此无需样本对生成内容进行规范与限制。

在问题生成完成后，问答生成模块进入回答生成阶段。高质量的问答数据集需满足3个基本要求：问题种类丰富多样、回答准确详实和问答对原文覆盖率高。回答生成阶段的主要挑战在于保证大模型回答质量的稳定性，具体体现在：(1)回答需严格依据原文信息，避免大模型幻觉问题；(2)回答需保持稳定输出长度，避免过于简短，确保涵盖问题所指目标的完整信息。为此，本文采用提示工程中的上下文学习方法：首先通过零样本方式生成问答对并进行人工筛选，构建小规模高质量问答及上下文样本集合；随后在多样本提示词中随机选取样本嵌入，以避免生成内容单一。在确定示例嵌入数量时，需兼顾大模型上下文窗口长度限制与长文本提示词导致的上下文遗忘问题。实验表明，嵌入5个或更多样本时，模型易遗忘提示词前段的约束，影响问答质量。由于该阶段输入(问题与上下文片段)通常超过600字符，加上输出回答后单个样本超过1 000字符，经尝试后选定每次嵌入3个样本，使提示词总长度不超过4 000字符，以达到最佳生成效果。问题生成及回答生成提示词示例如表2所示。

表2 生成阶段提示词示例

Table 2 Example of prompts for the generation phase

问题生成提示词：

你的任务是从〈文本〉生成适合作为问答对数据集的question，并输出能够解答该问题的context。以下是[任务要求]:

[你的输出包含question和context两部分。你应该保证问题只跟context中的信息有关。

question中只能包含一个问题。问题应该简洁并且有概括性。问题应该是航空发动机领域的专业性问题。context内应该是完整的句子。

你应该忽略〈文本〉提及的任何图片信息，context部分不能出现类似‘如图’的描述。你应该忽略〈文本〉提及的任何公式信息。

如果〈文本〉中存在你无法理解的公式或数字信息，则直接输出"无法提取"。

你的输出应该使用标准JSON格式数据。]

〈文本〉:{{context}}

回答生成提示词：

你是一位问答对数据处理专家。你需要帮助我生成高质量问答对数据集。

我会输入question和context，你的任务是根据context回答该question，输出answer。以下是[任务要求]:

[你的输出只包含answer部分，

answer部分应该非常详细而准确。answer部分应该分点进行回答。answer不能提到图片信息。

你应该充分理解context的信息，通过推理归纳成航空领域的知识后，再输出answer。

如果根据context无法准确回答问题，你应该直接输出“无法回答”。

使用标准JSON格式。]

示例1:〈{question:[eg_q],context:[eg_c]},{answer:[eg_a]}〉

⋮

示例3:〈{question:[eg_q],context:[eg_c]},{answer:[eg_a]}〉

human:{question:{{input_q}},context:{{input_c}}}

assistant:

在航空发动机领域问答数据集的构建中,本文采用DeepSeek API调用deepseek-V2.5模型进行生成任务。DeepSeek系列[27]是采用了多头潜在注意力结构和稀疏混合专家模型技术的国产开源大模型。在中文测试集的对战式评测中,deepseek-V2.5对ChatGPT-4o-latest的胜率达43%,表明其在中文对话场景中具备接近SOTA模型的通用能力,实验结果验证了deepseek-V2.5在中文对话文本生成任务中的优异表现,且其调用成本显著低于ChatGPT系列。

2.4 数据过滤

经过数据生成阶段,得到未过滤的问答数据集$\boldsymbol{H}=\{(q_1,a_1,c_1),(q_2,a_2,c_2),\cdots,(q_n,a_n,c_n)\}$,每条数据由问题、回答和上下文3部分组成。本文旨在构建高质量的开放型闭卷式问答数据集,在后续模型实验中转换为指令微调数据集时将剔除上下文部分。因此,数据过滤阶段重点关注问答数据可能存在的两类缺陷:(1)忠诚度低。受幻觉问题影响,大模型可能生成超出上下文范围或事实错误的虚假信息,导致问答数据的事实准确性不足;(2)语义质量低下。由于航空发动机领域语料稀疏,大模型预训练的通用能力难以覆盖领域专业知识需求,可能导致生成回答时出现一致性或相关性等语义错误。针对上述问题,本文提出基于词向量嵌入余弦相似度的忠实度评估与基于大模型评估的语义质量评估相结合的方法,共同作用于过滤阶段,以确保最终数据集的高质量。

2.4.1 忠实度评估

本文提出了一种基于语义相似度的问答数据忠实度评估方法。该方法通过计算回答中与原文语义相似的句子比例来评估忠实度。与传统的字符级相似度计算方法[28](如编辑距离、Jaccard相似度等)不同,本文采用基于双向编码器(Bidirectional encoder representations from Transformers, BERTs)结构的句子嵌入模型[29]结合余弦相似度的策略,以捕捉更深层次的语义信息。具体而言,首先利用BERT的预训练语言模型将文本转换为上下文相关的向量表示,该模型通过双向Transformer架构捕获句子级的语义表征,能够更好地处理一词多义和长距离依赖问题。在此基础上,采用余弦相似度计算向量间的语义相似度,其值域范围为[−1,1],值越大表明语义相似度越高。实验采用Google预训练的双向编码句子嵌入模型LaBSE[30]进行句子向量化处理,最终获得忠实度评分。

本阶段过程可形式化描述如下:给定回答与上下文的数据对$(a_i,\ c_i)$及嵌入模型M_b,通过嵌入模型得到其向量表示$\boldsymbol{v}_a=M_b(a_i)$,同时对上下文进行嵌入得到向量集合$\boldsymbol{V}_c$。随后对$\boldsymbol{v}_a$与$\boldsymbol{V}_c$中各元素$\boldsymbol{v}_c$计算余弦相似度,有

$$\boldsymbol{v}_{\cos}=\frac{\boldsymbol{v}_a\cdot\boldsymbol{v}_c}{\|\boldsymbol{v}_a\|\times\|\boldsymbol{v}_c\|} \tag{1}$$

得到a_i中每个句子$\boldsymbol{s}_i$对应的相似度值$\boldsymbol{v}_{\cos}$。设定有相似度阈值$\boldsymbol{T}_{\cos}$,若$\boldsymbol{v}_{\cos}>\boldsymbol{T}_{\cos}$则将$\boldsymbol{s}_i$归类为相关句子$\boldsymbol{S}_{\mathrm{rel}}$,否则归类为无关句子$\boldsymbol{S}_{\mathrm{irr}}$。最终得到该回答$a_i$对应相关句子集合$\boldsymbol{S}_{\mathrm{rel}}$和无关句子集合$\boldsymbol{S}_{\mathrm{irr}}$,从而计算该回答$a_i$的忠实度得分$\boldsymbol{f}$为

$$f=\frac{|\boldsymbol{S}_{\mathrm{rel}}|}{|\boldsymbol{S}_{\mathrm{rel}}|+|\boldsymbol{S}_{\mathrm{irr}}|} \tag{2}$$

式中$|\boldsymbol{S}_{\mathrm{rel}}|$和$|\boldsymbol{S}_{\mathrm{irr}}|$分别表示相关句子和无关句子的数量。得分越高,表明回答段落中与原文相关的句子比例越大。基于过滤前的26 000余条数据,图4展示了忠实度得分的分布情况。

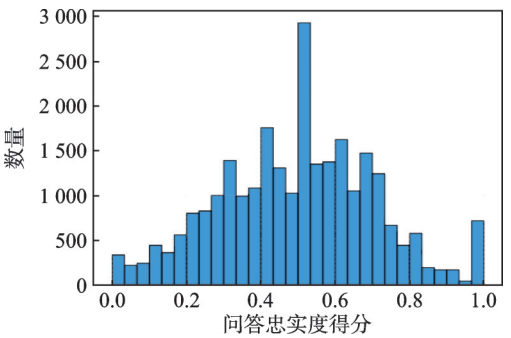


图4 过滤前问答数据忠实度分布

Fig.4 Distribution of faithfulness in question-answering data before filtering

2.4.2 语义质量评估

本文针对数据集设置了3个大模型语义评估指标:(1)相关性。衡量回答与问题的匹配程度,反映模型对问题的理解能力;(2)合理性。评估回答文本的逻辑一致性,确保内容前后连贯;(3)可靠性。检验回答是否基于上下文信息生成,避免凭空捏造。由于问答生成阶段采用输入优先策略,确保了问题均基于上下文提出,无需额外知识解答。其中可靠性指标作为对回答忠实度的补充评估,与基于原文相似度的忠诚度评估结果共同保障过滤后数据的高事实准确率。

本文在语义评估设计中引入了可解释性推理方法[31],要求评估大模型在输出评估结果时不仅提供得分,还需附上推理依据,以增强其归纳推理能力及结果的可解释性。基于过滤前的26 000余条数据,图5展示了3种语义评估指标(相关性、合理性、可靠性)的通过数据与未通过数据的对比结果。

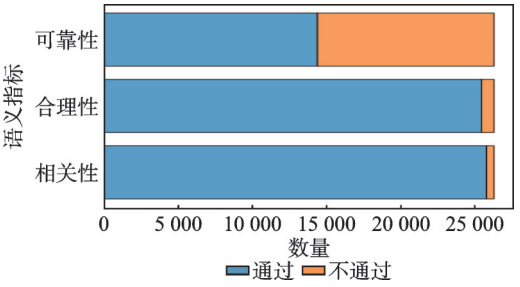


图5 3种语义质量指标评估结果统计

Fig.5 Statistics of three semantic quality metrics evaluation

生成问答对的质量缺陷主要集中于可靠性方面,这与大模型普遍存在的幻觉问题密切相关。尽管在数据生成阶段采用了多样本提示和分阶段生成策略以缓解幻觉问题,但仍有大量生成数据因事实可靠性不足而被过滤。相比之下,相关性与合理性未通过的样本较少,主要原因如下:(1)输入优先策略与分步生成机制有效避免了答非所问现象,表明通用大模型在问答任务中具备较强的对话相关性保持能力;(2)合理性评估涉及复杂的推理验证,且难以通过提示词全面描述否定条件,导致大模型倾向于给出乐观判断,从而降低了否定结果的准确性。

2.4.3 过滤策略

经过忠实度评估和语义质量评估,原始问答数据集$\boldsymbol{H}$在忠实度维度和语义质量维度上获得了不同评分。针对忠实度维度,基于图4所示的统计分布,采用回归树[32]中的最优分割目标函数确定忠实度过滤阈值$T_{\mathrm{f}}=0.537$,从而划分出忠实度通过集。具体而言,阈值T_{f}通过最小化高分集和低分集内数据点与各自均值的平方差之和确定,如式(3)所示,以确保两集合间的差异最大化。

$$T_{\mathrm{f}}=\arg\min_{T}\sum_{j\in F_{\mathrm{low}}}\left(H_F(x_j)-\hat{H}_{\mathrm{low}}\right)^2+\sum_{j\in F_{\mathrm{high}}}\left(H_F(x_j)-\hat{H}_{\mathrm{high}}\right)^2 \tag{3}$$

$$F_{\mathrm{low}}=\left\{j:H_F(x_j)<T\right\},\ \hat{H}_{\mathrm{low}}=\frac{1}{|F_{\mathrm{low}}|}\sum_{j\in F_{\mathrm{low}}}H_F(x_j) \tag{4}$$

$$F_{\mathrm{high}}=\left\{j:H_F(x_j)>T\right\},\ \hat{H}_{\mathrm{high}}=\frac{1}{|F_{\mathrm{high}}|}\sum_{j\in F_{\mathrm{high}}}H_F(x_j) \tag{5}$$

式中:$H_F(x_j)$表示原始问答数据集H中的一条样本x_j的忠实度得分。

对于语义质量维度,本文设定仅当3个子维度(相关性、合理性和可靠性)均通过评估的元素方可划入语义质量通过集。最终,通过两个维度过滤得到高质量问答数据集,并采用留出法以约3:1的比例随机划分训练集和测试集用于实验。表3展示了过滤完成后原始数据集的划分数量及统计结果。

表3 过滤阶段各集合数据数量

Table 3 Data quantity in each set during the filtering phase

集合类型	数据数量
原始数据集	26 245
忠实度通过集	13 963
语义质量通过集	14 325
EngineQA训练集	6 820
EngineQA测试集	2 000

3 实验与分析

为评估本文框架构建的问答数据集对航空发动机领域知识问答任务的表现提升效果,从高质量数据集EngineQA中随机划分训练集与测试集,并选取多个具备中文对话能力的

开源大模型。通过对其中一个大模型在训练集上进行微调,并在测试集上对比微调前后模型的对话表现,验证了该框架生成的数据集在微调后显著提升了大模型在航空发动机领域的知识问答能力。

3.1 实验设置

本文实验选取了4个开源通用模型:基于Llama2-7B-chat模型并经过中文数据集对齐的Llama2-Chinese-7B-chat模型、通过直接偏好优化(Direct preference optimization,DPO)与结构优化改进的ChatGLM2-6B模型、经过高质量语料训练优化的Baichuan2-7B-chat模型,以及基于Qwen1.5系列继续预训练与DPO优化的Qwen2-7B-Instruct模型。为确保微调模型性能验证的有效性,所有对比模型均选用同尺寸且在初代模型基础上优化的二代对话版本,并采用零样本提示词,使测试结果更贴近真实问答环境中的泛化性能。

在指令微调阶段,本文采用LoRA微调[33]方法对Qwen2-7B-Instruct模型进行微调,使用6 820条训练数据,得到航空发动机领域问答大模型Qwen2-7B-Engine。微调过程在8×A5000服务器上进行,超参数设置参考千条级别指令数据集的常规配置,如表4所示。

为全面评估模型回答质量,本文实验不仅采用基于n-gram的常用指标(BLEU、ROUGE和METEOR),还引入了基于BERT模型的BERTScore指标。

表4 LoRA微调超参数

Table 4 Hyperparameters for LoRA fine-tuning

超参数名	数值
单设备训练批次	4
梯度累积步数	4
训练轮次	5
学习率	0.000 1
LoRA秩	8
LoRA缩放因子	32
LoRA Dropout率	0.1

3.2 结果分析

不同模型在测试集上的表现如表5所示。结果显示,经LoRA微调后的Qwen2-Engine模型在各指标上均较同尺寸通用模型及微调前模型提升5%以上。该结果表明,基于本文问答数据生成方法构建的航空发动机领域问答数据集,能够通过指令微调显著提升大模型在该领域问答任务中的事实准确性,验证了该方法生成的数据在垂直领域问答任务微调中的有效性。

表5 不同模型在测试集上的事实准确性评估

Table 5 Evaluation of fact accuracy across different models on test set

模型	参数量/B	BLEU	ROUGE	METEOR	BERTScore
Llama2-Chat	7	18.17	28.74	23.04	70.77
Qwen2-Instruct	7	17.15	31.37	33.87	71.80
Baichuan2-Chat	7	17.89	33.27	24.84	71.56
ChatGlm2	6	15.61	30.93	21.26	70.84
Qwen2-Engine	7	28.32	41.88	36.82	76.46

为验证该数据集在模型微调任务中不仅能编辑领域知识至参数化记忆,还能提升生成回答的语义质量与稳定性,本文采用更大参数量的开源模型Qwen2.5-72B-Instruct,对各模型生成的问答对数据进行准确性、完整性和合理性3个维度的语义质量评估。评估结果如图7所示。微调前的4个通用模型中,Baichuan2-Chat在3个维度上得分最高;而微调后的

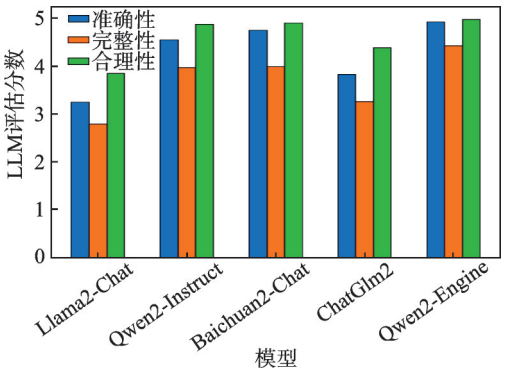


图7 不同模型在测试集上的语义质量评估

Fig.7 Semantic quality evaluation of different models on the test set

Qwen2-Engine在准确性、完整性和合理性上均优于微调前模型，并超越Baichuan2-Chat。结果表明，该数据集用于微调能显著提升模型生成回答的语义质量与稳定性，使其更贴合问答场景需求。

3.3 案例研究

为了直观展示本文所构建的高质量航空发动机领域数据集对大语言模型问答能力的提升效果，本节选取了一个典型案例进行分析。具体地，考察了模型在经过该数据集进行指令微调前后，针对航空发动机领域中的一个具体专业问题："如何实现涡轮机不平衡振动补偿？"给出的回答差异。微调前的模型是Qwen2-7B-Instruct基础模型，微调后的模型则是在此基础模型上利用本文方法构建的航空发动机专业数据集进行指令微调得到的Qwen2-Engine模型。表6详细对比了模型微调前后对该问题的具体回答内容及Token消耗量。微调前的模型针对该问题提供了一个相对较长（消耗504 Tokens）的回答。该回答列举了诸如设计改进、制造精度、热处理和故障诊断等多种与减少或处理涡轮机振动相关的方法。虽然内容具有一定的相关性，但其涵盖范围较广，并未能精准聚焦于"不平衡振动补偿"这一特定技术目标，回答的逻辑结构也显得较为分散，未能体现出领域内的专业知识体系。

表6 模型微调前后问答能力对比

Table 6 Comparison of model question answering capability before and after fine-tuning

问题	微调前模型（消耗504 Tokens）	微调后模型（消耗304 Tokens）
如何实现涡轮机不平衡振动补偿	在航空领域中，涡轮机（通常指涡轮喷气发动机或涡轮风扇发动机中的涡轮部分）的不平衡振动是一个严重的问题，实现涡轮机的不平衡振动补偿通常涉及以下几个方面： 1. 设计改进：叶片设计。优化叶片的几何形状和材料，减少叶片质量分布的不均匀性。例如，使用轻质合金或复合材料减轻重量，同时确保叶片的强度和刚度。平衡片。在叶片的适当位置添加平衡片，以调整叶片的质量分布，从而减少不平衡量。 2. 制造精度：精密加工。采用高精度的加工设备和技术来制造涡轮叶片和转子，减少制造过程中的误差，提高部件的尺寸精度和表面光洁度。热处理：通过合理的热处理工艺，改善材料的力学性能，进一步减小由于热膨胀等因素引起的不平衡。 ⋮ 5. 故障诊断：利用先进的诊断技术，如振动分析、油液分析等，预测可能的故障并采取预防措施。 通过上述方法的综合应用，可以有效地减少涡轮机的不平衡振动，提高发动机的可靠性和运行效率。在实际应用中，这些技术往往结合使用，以达到最佳的补偿效果。	在航空领域中，实现涡轮机的不平衡振动补偿主要通过以下几种方法： 1. 转子结构优化设计。通过对转子结构进行优化设计，减少不平衡量和不平衡响应。例如，采用特定的支承方案和轴承设计，以提高转子系统的稳定性和抗振性能。 2. 转子动力学特性分析。利用转子动力学特性分析，确定转子的不平衡响应和振动特征。通过仿真模型模拟转子在不同工况下的振动情况，为实际工作中的振动补偿提供依据。 ⋮ 5. 实时监测与补偿。通过实时监测转子系统的振动状态，根据监测数据调整转子的支承位置或施加外力补偿，以消除或减小不平衡振动。 综上所述，实现涡轮机的不平衡振动补偿涉及转子结构优化、动力学特性分析、不平衡响应分析以及实时监测与补偿等多个环节，通过综合运用这些方法，可以有效降低涡轮机的振动水平，提高其稳定性和可靠性。

相比之下，经过本文方法构建的高质量数据集进行指令微调后的模型，给出了一个更为精炼（仅消耗304 Tokens）且指向性更强的回答，该回答清晰地阐述了实现涡轮机不平衡振动补偿的关键技术途径，包括转子结构优化设计、转子动力学特性分析以及实时监测与补偿等。这些方法更加准确地对应了问题的核心需求，且回答的专业性和逻辑性显著增强，体现了模型在经过该数据集微调后，能够更好地理解和解答航空发动机领域的具体问题。

4 当前问题与未来研究方向

高质量的问答数据集是构建实用可靠对话大模型应用的基础,然而在航空研制等知识密集且专有性强的垂直领域,由于人工构建难度大、成本高、对标注人员专业知识要求高等因素,相关研究与应用一直受限于数据集的匮乏。本文提出了一种问答数据集生成方法,成功构建了航空发动机领域的高质量问答数据集EngineQA,并通过实验验证了大模型能够通过合理的提示工程设计生成垂直领域数据集。本文的研究工作为垂直领域数据集的生成以及开放型闭卷式问答数据质量的评估,提供了有益的参考和借鉴。本文工作仍存在一定局限性。未来研究可从以下几个方面展开,以进一步推动数据生成与数据质量评估领域的发展。

(1) 生成模型的选择。本文工作聚焦于数据生成与过滤策略的选择和尝试,在生成模型的选择上只采用了在中文对话能力上接近SOTA水平的国产模型Deepseek-V2.5。随着后续大模型领域研究发展以及大模型的能力不断提升,在未来工作中可以选择性能更好、长文本能力更强的模型,以进一步提高生成质量从而减少过滤导致的信息流失。

(2) 大模型幻觉缓解技术。尽管本文通过多样本提示词、输入优先策略等方法缓解了大模型幻觉问题,但在过滤阶段仍因事实可靠性不足而筛除大量数据。未来可结合学术界对幻觉问题的最新研究成果,采用更先进的缓解技术,进一步提升生成数据的可靠性。

(3) 多模态数据集的构建。本文仅构建了文本形式的问答数据集,而在航空发动机等复杂工程领域,图片、表格等多模态数据同样具有重要价值。未来可探索构建多模态数据集,以支持故障检测等任务,提升大模型在多模态场景下的能力,这将是垂直领域大模型研究的重要方向之一。

5 结束语

本文提出了一种基于大语言模型的高质量问答数据自动生成方法,通过上下文学习与输入优先策略,显著提升了生成数据的稳定性与一致性。在数据过滤环节,本文设计了一种开放式问答数据质量的自动评估机制,该机制结合了基于原文相似度的忠实度评估与基于大模型的语义质量评估,有效过滤了生成数据中的幻觉问题与语义错误,确保了数据集的事实可靠性与语义质量。基于该方法,本文构建了航空发动机领域的开放式问答数据集EngineQA,并通过与多个通用大模型的测试对比,验证了该数据集在领域知识问答能力提升方面的有效性,同时证明了其适用于航空发动机领域专家问答大模型的指令微调任务。本文的研究成果不仅为航空发动机领域的大模型应用提供了数据基础,也为其他垂直领域的数据集构建与质量评估提供了参考与借鉴。

参考文献:

[1] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[2] TOUVRON H, LAVRIL T, IZACARD G, et al. LLAMA: Open and efficient foundation language models[EB/OL].(2023-02-27)[2025-01-08]. https://arxiv.org/abs/2302.13971.

[3] YANG A, YANG B, HUI B, et al. Qwen2 technical report[EB/OL]. (2024-09-10) [2025-01-08]. https://arxiv. org/abs/2407.10671.

[4] 夏润泽,李丕绩. ChatGPT大模型技术发展与应用[J].数据采集与处理,2023,38(5): 1017-1034.
XIA Runze, LI Piji. Large language model ChatGPT: Evolution and application[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 1017-1034.

[5] 陈浩泷,陈罕之,韩凯峰,等. 垂直领域大模型的定制化:理论基础与关键技术[J].数据采集与处理,2024,39(3): 524-546.
CHEN Haolong, CHEN Hanzhi, HAN Kaifeng, et al. Domain-specific foundation-model customization: Theoretical

foundation and key technology[J]. Journal of Data Acquisition and Processing, 2024, 39(3): 524-546.

[6] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[7] FAN W, DING Y, NING L, et al. A survey on rag meeting LLMs: Towards retrieval-augmented large language models[C]// Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2024: 6491-6501.

[8] CHUNG H W, HOU L, LONGPRE S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.

[9] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[10] LIU N F, LIN K, HEWITT J, et al. Lost in the middle: How language models use long contexts[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173.

[11] MITKOV R. Computer-aided generation of multiple-choice tests[C]//Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 17-22.

[12] HEILMAN M, SMITH N A. Good question! statistical ranking for question generation[C]//Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California: Association for Computational Linguistics, 2010: 609-617.

[13] LEE D B, LEE S, JEONG W T, et al. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2020: 208-224.

[14] VIRANI A, YADAV R, SONAWANE P, et al. Automatic question answer generation using T5 and NLP[C]//Proceedings of the 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS). [S.l.]: IEEE, 2023: 1667-1673.

[15] SAADANY H, ORASAN C. BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text[C]//Proceedings of the Translation and Interpreting Technology Online Conference. [S.l.]: INCOMA Ltd., 2021: 48-56.

[16] SONG Y, MIRET S, ZHANG H, et al. HoneyBee: Progressive instruction finetuning of large language models for materials science[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023: 5724-5739.

[17] YUE X, WANG B, CHEN Z, et al. Automatic evaluation of attribution by large language models[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023: 4615-4635.

[18] WAN Y, AJITH A, LIU Y, et al. SciQAG: A framework for auto-generated scientific question answering dataset with fine-grained evaluation[EB/OL].(2024-07-10)[2025-01-08]. https://arxiv.org/abs/2405.09939.

[19] JIN Q, DHINGRA B, LIU Z, et al. PubMedQA: A dataset for biomedical research question answering[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hongkong, China: Association for Computational Linguistics, 2019: 2567-2577.

[20] GUHA N, NYARKO J, HO D, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2024, 36: 44123-44279.

[21] JIN D, PAN E, OUFATTOLE N, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams[J]. Applied Sciences, 2021, 11(14): 6421.

[22] ZIEGLER I, KÖKSAL A, ELLIOTT D, et al. CRAFT your dataset: Task-specific synthetic dataset generation through corpus retrieval and augmentation[EB/OL].(2024-09-03)[2025-01-08]. https://arxiv.org/abs/2409.02098.

[23] DONG Q, LI L, DAI D, et al. A survey on in-context learning[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, FL, USA: Association for Computational Linguistics, 2024: 1107-1128.

[24] WANG Y, KORDI Y, MISHRA S, et al. Self-Instruct: Aligning language models with self-generated instructions[C]//

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 13484-13508.

[25] MA Y, YU D, WU T, et al. PaddlePaddle: An open-source deep learning platform from industrial practice[J]. Frontiers of Data and Domputing, 2019, 1(1): 105-115.

[26] MANKU G S, JAIN A, DAS SARMA A. Detecting near-duplicates for web crawling[C]//Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada: ACM, 2007: 141-150.

[27] LIU A, FENG B, WANG B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model[EB/OL].(2024-07-19)[2025-01-08]. https://arxiv.org/abs/2405.04434.

[28] 鲍静益,于佳卉,徐宁,等.问答系统命名实体识别改进方法研究[J].数据采集与处理,2020,35(5): 930-941.
BAO Jingyi, YU Jiahui, XU Ning, et al. Research on the improved method of named entity recognition in Q &. A system[J]. Journal of Data Acquisition and Processing, 2020, 35(5): 930-941.

[29] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hongkong, China: Association for Computational Linguistics, 2019.

[30] FENG F, YANG Y, CER D, et al. Language-agnostic BERT sentence embedding[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 878-891.

[31] JIE Y W, SATAPATHY R, GOH R, et al. How interpretable are reasoning explanations from prompting large language models?[C]//Proceedings of Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City, Mexico: Association for Computational Linguistics, 2024: 2148-2164.

[32] LOH W Y. Classification and regression trees[J]. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 2011, 1(1): 14-23.

[33] HU E J, WALLIS P, ALLEN-ZHU Z, et al. LoRA: Low-rank adaptation of large language models[EB/OL].(2021-06-17)[2025-01-08]. https://arxiv.org/abs/2106.09685.

作者简介:

邹冠沄(1999-),男,硕士研究生,研究方向:自然语言处理、大模型,E-mail: zouguanyun@nuaa.edu.cn。

王存俊(1995-),男,工程师,研究方向:民机设计知识智能化,领域大模型应用,E-mail: wangcunjun@comac.cc。

孔寅豪(1998-),男,助理工程师,研究方向:领域大模型评测,民机设计知识语料构建,E-mail:kongyinhao@comac.cc。

马小庆(1983-),男,研究员,研究方向:知识图谱、大语言模型等,E-mail: maxiaoqing@comac.cc。

李丕绩(1986-),**通信作者**,男,教授,博士生导师,研究方向:自然语言处理、大模型、多模态等,E-mail:pjli@nuaa.edu.cn。

(编辑:刘彦东)