

DOI:

基于大语言模型的结构件加工工艺推荐方法研究*

郑小虎^{1,2} 陈宏博³ 何方舟⁴

- (1. 东华大学人工智能研究院 上海 201620;
2. 上海工业大数据与智能系统工程技术研究中心 上海 201620;
3. 东华大学机械工程学院 上海 201620;
4. 上海飞机制造有限公司 上海 201324)

摘要: 在复杂结构件数控编程过程中, 由于机械加工工艺知识的多源异构及知识之间的关联复杂性, 导致加工工艺知识重用困难。提出一种基于大语言模型的结构件加工工艺知识推荐方法, 通过大语言模型的选择和微调, 建立结构件加工工艺知识推荐的垂直领域模型, 效果评估表明该模型可以针对具体零件特征推荐出相应的加工工艺。为解决模型无法获取最新专业知识且加工工艺推荐准确度低的问题, 利用 LangChain 应用框架结合知识库对领域模型进行知识检索增强并构建工艺知识问答系统, 通过相应指标评价, 实现了该问答系统在原有领域模型的基础上 F1 值提升 0.026, 加工工艺推荐准确度在 90% 以上。在航空结构件数控编程的工艺决策应用中, 该方法针对零件特征实现了相应工艺知识的推荐, 与未使用本文方法的自动数控编程系统相比, 框体类结构件数控代码生成效率具有一定提升, 对提高数控编程人员的决策效率具有重要意义。

关键词: 大语言模型; 加工工艺; 知识推荐; 快速编程

中图分类号: TH16

Research on Machining Process Recommendation Method of Structural Parts Based on Large Language Models

ZHENG Xiaohu^{1,2} CHEN Hongbo³ HE Fangzhou⁴

- (1. Institute of Artificial Intelligence, Donghua University, Shanghai 201620;
2. Shanghai Industrial Big Data and Intelligent Systems Engineering Technology Center, Shanghai 201620;
3. College of Mechanical Engineering, Donghua University, Shanghai 201620;
4. Shanghai Aircraft Manufacturing Company Limited, Shanghai 201324)

Abstract: In the process of numerical control programming for complex structural components, the difficulty in reusing machining process knowledge arises due to the heterogeneity of knowledge sources and the complexity of interconnections between knowledge. A knowledge recommendation method for structural parts machining process based on a large language model is proposed. By selecting and fine-tuning the large language model, a vertical domain model of machining process knowledge recommendation for structural parts is established. The evaluation results indicate that the model can recommend corresponding machining processes based on specific part features. To solve the problem of the model not being able to obtain the latest professional knowledge and the low accuracy of machining process recommendations, the LangChain application framework combined with a knowledge base is used to enhance the knowledge retrieval of the domain model and construct a process knowledge question answering system. Through corresponding indicator evaluation, the F1 value of the question answering system improves by 0.026 on the basis of the original domain model, and the accuracy of machining process recommendations is above 90%. In the process decision-making application of CNC programming for aviation structural components, this method recommends corresponding process knowledge based on part

* 国家自然科学基金青年科学基金(51805079)、中央高校基本科研业务费专项资金(2232021D-15)资助项目。

features. Compared with the automatic CNC programming system that does not use the method in this article, the efficiency of generating CNC codes for frame type structural components improves to a certain extent, which is of great significance for improving the decision-making efficiency of CNC programmers.

Key words: large language models; machining process; knowledge recommendation; rapid NC programming

0 前言

航空结构件作为飞机中支撑作用占比最大的零件,其生产周期直接影响机型产能^[1]。由于目前缺乏有效的飞机结构件数控加工编程知识共享和重用工具,导致企业内部积累的大量数控工艺知识无法重复利用。目前基于三维模型的工艺知识重用方法主要依赖文本或内容检索,通过加工意图识别可提高工艺知识重用效能^[2],而设计效率难以保证,如何将丰富的数据知识高效地重用与推荐是当前研究的难点。新兴的大语言模型作为一种知识调用技术有望为机械加工工艺知识的表达与推荐提供有效的解决途径。

在零件加工过程中,针对工艺知识进行高效重用的问题,国内外学者主要从数据挖掘技术与工艺知识的表达方法上进行突破。在工艺知识挖掘方面,ZHOU 等^[3]以零件特征为重点,建立工艺路线的模糊相似度矩阵,通过粒度计算得到典型的工艺序列;罗滨鸿等^[4]根据网线编织工艺数据的特点对 Apriori 算法进行了改进,实现了网线编制车间里工艺数据的关联规则挖掘;阳树梅等^[5]将零件几何信息进行数值化表达,选用聚类算法结合主成分分析等降维方式对历史工艺数据进行聚类分析以获得典型工艺模板。在知识表达方面,ZHOU 等^[6]利用车间资源知识图谱挖掘隐式资源信息,进行知识图谱的更新,实现知识的有效表达;ZHENG 等^[7]利用多智能体强化学习算法,在面向工业知识图谱过程中通过自我配置、优化、适应、调整,达到自我认知水平,实现多语义下的自我配置解决方案搜索和任务的分解。ZHOU 等^[8]以因果理论为基础,对航天产品制造质量缺陷中的数据进行因果关系图构建,并基于此开发了工业因果知识增强的大语言模型,提高了产品质量问题分析的专业性与可靠性。在以上的研究中,通过基于模型的工艺知识挖掘方法可以获取企业生产制造中的隐性知识。然而,工艺数据复杂且分散,来源于不同系统如 CAD、MES 和 ERP,包含结构化数据(如加工参数、零件特征)和非结构

化数据(如工艺说明、专家经验),加之数据质量问题如冗余、错误和缺失,使得传统数据挖掘方法难以全面获取完整的工艺知识。尽管知识图谱技术能够形象化表达知识,但受限于数据不足和技术实现复杂,导致实施难度大且可扩展性差。相关学者通过结合知识图谱与大语言模型技术的特点,挖掘数据间的因果知识,为知识的表达与高效推荐提供有益指导。

大语言模型作为一种基于深度学习技术的自然语言处理模型,它可以针对大量文本数据进行训练学习,自动识别和理解文本中的语义信息和知识,并进行高效推荐^[9-10]。随着模型参数量的增长,在消费级硬件上对模型进行训练或者全部参数的微调变得不可行。KOJIMA 等^[11]通过构建思维链式提示词,引导模型进行多步推理,证明大语言模型具有基本的零样本微调能力。近年来参数高效微调方法的出现,使得在模型微调过程中仅改变少量训练参数即可达到模型全参数微调的效果。为增强大语言模型在垂直领域的应用能力并降低训练成本,学术界提出了 P-Tuning v2、LoRA、QLoRA 等参数高效的模型微调方法^[12-14]。由于大语言模型在实际应用过程中无法获取最新专业知识,张鹤译等^[15]利用 LangChain 框架结合大语言模型与专业知识库提出基于提示学习的问答系统范式,增强了模型回答的专业性。赵鑫等^[16]针对已有信息检索架构的改进与大语言模型的融合方式,探讨了大语言模型时代下的信息检索发展趋势,为模型的知识获取提供技术手段。随着大语言模型研究的不断推进,模型微调技术与结合知识库的知识增强方法对机械制造领域工艺知识的重用与推荐具有借鉴意义。大语言模型在医疗、金融、电力等专业领域相继得到应用^[17-19],但在机械工艺推荐领域尚未见比较成熟的应用案例。

本文针对航空结构件数控编程过程中工艺知识的重用与推荐问题,引入大语言模型技术,提出一种基于大语言模型的结构件加工工艺知识推荐方法,通过构建机械加工工艺推荐领域模型,搭建领域模型与向量数据库的融合问答系统,经航空结构

件数控编程过程中的案例应用,验证了所提方法的有效性与高效性。

1 工艺知识推荐方法框架

针对航空结构件数控编程过程中存在的效率低、企业内部工艺知识无法高效重用与推荐的问题,为了提高工艺加工人员的决策效率,设计了图 1 所示的基于大语言模型的工艺知识推荐系统框架。该系统主要由机械加工领域基座模型的选择、加工工艺知识向量数据库的创建、基座模型的微调和工艺知识问答推荐系统的构建四个主要部分构成。

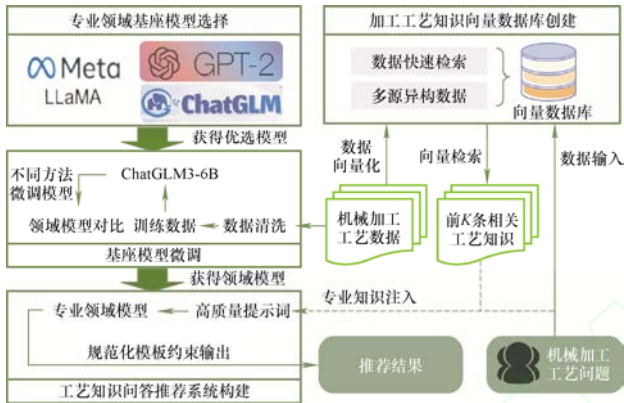


图 1 工艺知识推荐系统应用框架

1.1 专业领域基座模型的选择

为适配机械加工工艺知识推荐且考虑模型应用可商业化的特点,根据大语言模型的参数规模及基础训练数据内容进行模型对比优选,国内外一些常用的开源模型如表 1 所示。

表 1 国内外主要开源大语言模型

名称	支持语言	参数规模	基础训练数据
LLaMA ^[20]	多语言	7B/13B/65B	约 1.4 万亿标记
GPT2 ^[21]	多语言	1.5B	约 40GB 文本数据
Alpaca ^[22]	英文	7B	约 5.2 万条问答指令数据
ChatGLM ^[23]	中文、英文	6B/12B/130B	约 1TB 中英比例为 1:1 的标记
Baichuan	中文、英文	7B/13B	约 1.2 万亿标记

在国外开源的模型中, Alpaca 模型是基于 LLaMA 模型进行微调得到的指令遵循模型,其特点在于小巧且易于复现,因此它非常适合用于学术研究,但 Alpaca 模型禁止任何商业用途。LLaMA 模型的训练数据范围非常广泛,其 13B 的模型性能可以达到 GPT3 的效果。至于 GPT2 模型,其诞生时间较早,模型参数量也相对较小,为研究者们探索大规模预训练语言模型的内部工作机制和性能提供

了有益的帮助。然而,这些模型都存在一些固有的局限性,包括处理中文能力较差,部署困难,无法获得关于最近的最新信息等。对于国内的开源大语言模型,主要有百川智能开发的 Baichuan 系列模型以及清华大学开源的 ChatGLM 系列模型。它们在预训练阶段都采用了大量的中文语料,从而能够更好地理解和生成中文文本,相对于 Baichuan 系列模型,ChatGLM 系列模型因其轻量化且提供专有微调方法^[12]更受开发人员青睐。

本研究针对机械加工领域中文知识进行推荐,考虑实验室计算资源适中及模型可商用特点,采用清华大学开源的 ChatGLM 系列模型作为基座模型,为保证模型实际应用效果,选择参数量适中且性能更优的第三代 ChatGLM3-6B 模型。

1.2 加工工艺知识向量数据库的创建

为提高模型的实际应用效果,提供其获取领域专业知识及应用案例的能力,引入向量数据库对模型进行知识检索增强,使得机械加工领域工艺知识及案例可以有效储存及重用,图 2 展示了工艺知识向量数据库的创建流程。

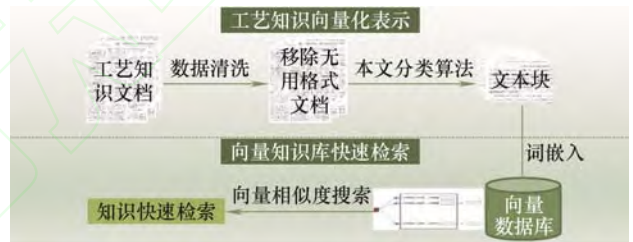


图 2 加工工艺知识向量数据库创建流程

加工工艺知识向量数据库的创建主要分为工艺知识的向量化表示与向量知识库快速检索两部分。首先收集和整理与工艺相关的各种资源,如工艺参数、工艺流程、操作方法、技术文献等文档共 2.48 GB。之后对工艺文档数据进行数据清洗,去除文档中无用数据及无关格式,然后利用文本切分算法根据中文语言写作特点进行段落划分使之保留原有的语义,提高向量化的准确性,通过词嵌入模型将中文数据转化为向量形式储存在向量数据库中,根据向量相似度查询算法实现高效的向量数据检索。

1.3 基座模型的微调

通常基座模型拥有的领域知识仅仅是训练过程中给予的大量通用知识,在垂直领域中模型的回答并不能满足专业需求,为挖掘基座模型拥有的隐性知识,激发其在机械加工领域工艺知识领域的知识推荐能力,进行模型的微调训练,使其更加适配领域任

务。由于高效微调的方法所需硬件资源较低,综合考虑实验室的硬件资源后,决定采用 P-Tuning v2、LoRA、QLoRA 三种主流参数高效的微调方法及冻结部分参数的 Freeze 方法对基座模型进行微调。模型微调的流程参考图 1 中基座模型微调部分,针对加工工艺知识离散化,工艺案例知识重用率低的特点,建立模型训练所需要的高质量问答数据,根据高效微调方法对基座模型进行微调对比选优,构建出工艺知识推荐任务的领域模型,通过模型评估实验,验证领域模型在工艺知识推荐任务上的可行性与专业性。模型微调部分具体细节及实验结果见第 2 节。

1.4 工艺知识问答推荐系统的构建

由于大语言模型在生成答案时常常表现出较高的自由度,这种自由性给模型输出带来了不确定性和不准确性。因此使用 LangChain^[24]开发框架,结合先前制作的向量数据库搭建起问答推荐系统以提升模型问答的可靠性和准确性并获取最新专业知识。LangChain 开发框架具备多种功能,可以协助开发人员构建定制的大语言模型对话产品。本研究采用 LangChain 框架,使用向量数据库来存储和管理领域知识,利用微调后的领域模型处理用户的输入,由工艺问题生成相应的输出结果,通过 LangChain 框架来管理和协调这些组件的交互,并提供一个统一的管理界面。工艺知识问答推荐系统搭建细节及评估结果见第 3 节。

2 加工工艺推荐领域模型构建方法

在机械加工工艺领域进行工艺知识推荐,由于加工任务的不同,工艺人员所需的工艺知识有所区别,为进一步提高模型知识推荐的精确性,需明确模型具体工作的场景以更好地适配工艺人员进行作业的任务。本研究结合航空结构件数控编程的任务需求进行相应的工艺知识推荐,通过高效的模型微调方法对比分析,根据选定的 ChatGLM3-6B 基座模型和自行构造的加工工艺数据集得到领域模型,根据试验结果的对比分析得到最终的加工工艺推荐领域模型。

2.1 数据制作和预处理

针对航空结构件数控编程任务中的工艺知识推荐需求,通过数控编程过程中的决策需求分析,发现工艺决策人员主要专注相关零件特征的加工工艺选择,且航空结构件中多框类结构件居多,如下图

3 所示。飞机结构件通常具有复杂的几何形状和严格的精度要求,这些结构件的加工特征也主要以槽腔类特征居多,并且主要依赖于数控铣削技术。面对大量待加工特征,数控加工人员重复工作量大、效率低且极易出错,为此针对多框类结构件特征进行领域数据集的制作。领域数据集主要考虑结构件加工过程中的大量工艺参数选择与加工方法的选用问题,突出零件加工过程中的机床、夹具、刀具等制造资源与进给、转速、切深、切宽等工艺参数的选择,以实现模型在飞机结构件关键参数选择的高可靠性推荐。为此,采用多源采样策略对相关已有的工艺加工案例进行梳理和分析,另外收集由加工特征及对应的加工方法组成的案例,部分案例见表 2。

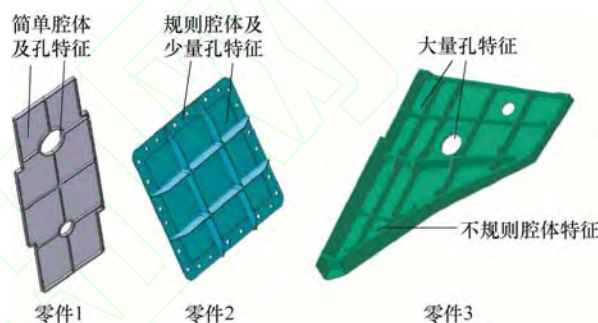


图 3 典型框体类结构件

表 2 机械加工零件特征对应加工方法案例

特征名称	长度/mm	宽度/mm	直径/mm	深度/mm	精度	粗糙度/ μm	加工方法
四边形腔	150	120	—	62	IT 8	6.3	粗铣 半精铣
通孔	—	—	50	88	IT 9	6.3	粗镗 半精镗
方形槽	200	50	—	20	IT 9	3.2	粗铣 半精铣 精铣
方形凸台	64	64	—	32	IT 9	6.3	粗铣 半精铣

在行业专家的指导下,为确保所收集到的数据具有较高的质量和准确性,对数据进行审查和校验,以剔除其中的错误和不准确的信息。通过对数据质量打分评估,确保得到的数据具有一定的有效性,同时对数据进行进一步的筛选和整理去掉数据中可能重复的部分。数据清洗步骤主要包括删除敏感信息、隐私信息和异常符号,并使用标点符号分隔文本。数据去重则采用 Minhash 算法以确保数据的唯一性。在人工手动整理完高质量数据问答对后,针对已有的高质量数据采用回译法与同义词替换等手段进行数据增强,提高数据集大小,数据集具体制作流程如图 4。

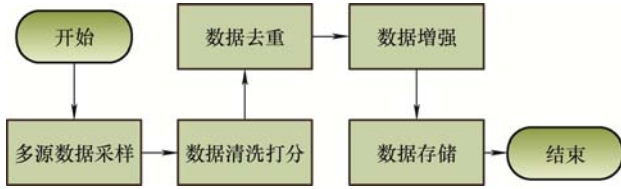


图4 模型微调数据制作流程

通过上述数据集制作流程，构建了基于特征的训练数据，数据总量为 5 000 条高质量特征工艺问答数据对，包含航空结构件数控编程任务中主要的加工特征，表 3 展示了 1 条针对四边形腔特征的示例训练数据。整体数据集由四边形腔、方形槽、圆形凸台、方形凸台、通孔、盲孔、外圆相关特征共 4 000 条以及其他工艺相关基础信息 1 000 条组成，该数据集可满足后续模型微调的需要。

表3 JSON 格式训练数据示例

键	值
指令	作为一名资深工艺加工工程师回答以下问题，通过给定加工特征与加工参数生成相应加工方法与工艺路线
输入	加工特征：四边形腔。加工参数：长度 120、宽度 100、直径 NULL、深度 50、精度 IT8、粗糙度 6.3 加工方法：粗铣-半精铣。加工工艺路线：选择 130 mm 长、110 mm 宽、60 mm 深度的铝合金工件材料。在 CNC 加工中心上进行定位和夹紧。使用直径 12 mm 的铣刀进行粗铣，对四边形腔进行粗加工，使其初步成形。更换铣刀，选择直径 4 mm 的铣刀进行半精铣，对四边形腔进行半精加工，以达到加工精度的要求
输出	

2.2 领域模型高效微调

为提高模型在数控编程任务中工艺知识的推荐能力，应用以上制作的数据集对模型进行高效微调。考虑试验环境及企业应用低成本需求，在微调方法上使用 P-Tuning v2、LoRA 和 QLoRA 三种主流的参数微调方法及冻结部分参数的 Freeze 微调方法进行对比实验分析，以期得到更优的领域模型。P-Tuning v2 方法在 transformer 每一层都加入了 Prompts tokens 作为输入，它使得模型可以有更多可学习的参数，实现足够参数高效。LoRA 方法对模型指定参数并行增加额外的低秩矩阵，仅训练额外增加的并行低秩矩阵的参数。在模型微调时，通过小的秩值选择，仅需训练很小的参数量，就能使模型获取较好的表现结果。根据大量消融试验，Hu 等^[13]指出当矩阵秩为 1、2、4、8、64 时，LoRA 方法的表现较为优异。QLoRA 是一种针对深度神经网络的低精度量化和微调技术，能够实现高保真的 4 位微调。Freeze 方法只需微调模型的高层特征，可以加快模型收敛、节约微调时间并最大程度保留大语言模型预训练阶段的大部分知识。通过对以上高效微调方法在数据集微调时的训练效果及最终领域模型能力的对比分析，确定适配机械加工领域工艺

知识推荐的高效微调方法并得到领域模型。

2.3 实验参数与指标设计

本文研究的具体实验环境配置如下：CPU，Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz；GPU，NVIDIA A40；显存：48；Python 版本，3.10.0；Cuda 版本，12.0；实验超参数设置如表 4 所示。为保证模型对于数据的充分利用并确保评价结果的客观性，本文通过 8：2 的方式划分训练集和测试集。为综合评估微调后的领域模型性能，选取 BLEU-N^[25]（含 N 个连续词元的双语评估替换分数）、ROUGE-N^[26]（含 N 个连续词元的召回率分数）、ROUGE-L^[26]（含最长公共子序列的召回率分数）指标对模型领域效果进行客观评估。由于自然语言的复杂性，为了更准确地评估领域模型的能力，使用通用数据集评估模型开放域能力以及将其与现有优异的 ChatGPT3.5 进行对比分析。

表4 微调实验超参数设置

实验超参数	实验方法			
	P-Tuning v2	LoRA	QLoRA	Freeze
最大输入序列长度	512	-	-	-
最大输出序列长度	512	-	-	-
每批次训练数据量大小	8	8	8	8
梯度累积步数	4	4	4	4
学习率	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
预序列长度	16	-	-	-
训练轮数	4	4	4	4
低秩矩阵的秩	-	8	8	-
缩放因子	-	16	16	-
量化位数	-	-	4	-
可训练层	-	-	-	-3

2.4 结果分析

通过使用先前制作的训练数据集，将 ChatGLM3-6B 模型使用主流的 LoRA、QLoRA 和 P-Tuning v2 三种参数高效微调方法，以及仅微调 Transformer 最后 3 层的全连接层并冻结其他所有参数的 Freeze 方法进行微调训练，训练损失结果如图 5 所示。

由上图可知，四种训练方法在实验条件整体一致的情况下，训练损失都有明显的下降。其中 LoRA 与 QLoRA 方法训练损失基本一致，主要是 QLoRA 方法相对 LoRA 方法只更改了训练过程中的参数精度，因此其训练损失符合预期。为进一步研究各种微调方法对模型的优化效果，对四种方法微调后的

模型进行评估, 选取此前制作的测试数据集进行实验, 模型评估指标为 BLEU-4、ROUGE-1、ROUGE-2、ROUGE-L, 具体结果如下表 5。

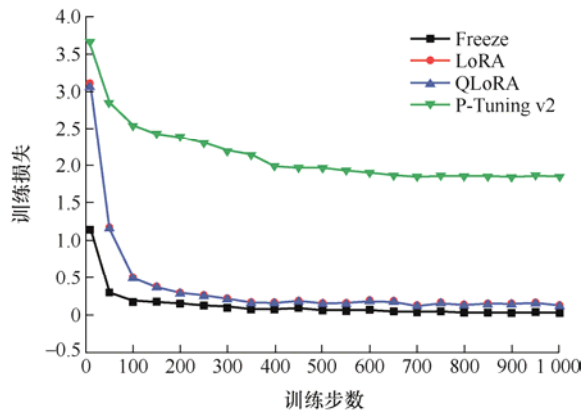


图 5 微调模型训练损失

表 5 微调前后模型效果评估

模型	BLEU4 (%)	ROUGE1 (%)	ROUGE2 (%)	ROUGEL (%)
LoRA 微调	54.32	72.23	57.42	66.85
QLoRA 微调	35.74	59.24	38.86	52.83
P-Tuning v2 微调	6.97	28.41	7.66	18.12
Freeze 微调	57.42	74.51	60.23	68.86
ChatGLM3-6b	4.31	26.89	4.71	15.24

由上表可以看出, 微调后的模型均比微调前的 ChatGLM3-6B 在领域数据上的认知效果有所提升, 说明通过领域数据集微调的方法可以提升大语言模型在领域知识问答上的效果。但 P-Tuning v2 方法微调后得到的模型在各项指标上均提升不高, 综合其微调损失下降情况分析, 该方法对领域知识的学习能力相对较弱。LoRA、Freeze 与 QLoRA 方法微调出的模型在领域数据上效果提升明显, 且 LoRA 与 Freeze 方法微调出的模型在各项指标上评价较优, 表明 LoRA 与 Freeze 方法更容易使模型适配数控编程中加工工艺知识的推荐任务。

为提高模型的领域适配性及实际应用效果, 进一步分析各种微调方法得到的领域模型表现差异, 统一四种微调方法的实验条件, 观察模型训练与推理的表现。四种方法训练与推理表现见下表 6, 从表中可以看出, 在含有约 62 亿参数的 ChatGLM3-6B 模型基础上, 除了 Freeze 微调方法外, 其他方法的可训练参数均降至 0.06%, 显著降低了 GPU 显存占用。QLoRA 方法通过将模型进行 4bit 量化, 虽然降低了训练显存, 但训练与推理速度明显下降。相较之下, LoRA 方法尽管显存占用略高, 但在训练和推理速度上表现优异。Freeze 方法的显存占用过高

且训练与推理速度远低于其他方法, 这严重限制了其应用。综合分析表明, LoRA 方法在可训练参数量、GPU 显存占用、训练速度和推理速度方面均优于 Freeze 方法, 同时在性能上也能接近 Freeze 的效果。因此, 选择 LoRA 方法进行 ChatGLM3-6B 模型的微调, 可以在节约资源的同时, 获得更优的领域适配效果。

表 6 微调模型训练与推理表现

训练方法	可训练参数量(%)	GPU 显存占用/GB	训练速度/(样本/s)	推理速度/(样本/s)
LoRA	0.06	16	7.67	1.68
QLoRA	0.06	8	6.51	1.40
P-Tuning v2	0.06	14	4.82	0.45
Freeze	8.09	28	1.58	0.08

在选取 LoRA 方法微调后的模型作为加工工艺推荐领域模型之后, 为评估该领域模型是否出现灾难性遗忘问题^[27], 保证模型在后续应用过程中具有足够的鲁棒性, 采取 CEVAL 和 CMMLU 中文大语言模型评估数据集与 MMLU 英文任务多样性评估数据集进行比较实验, 评估模型微调前后在以上三种数据集上的表现, 评估结果如图 6 所示。

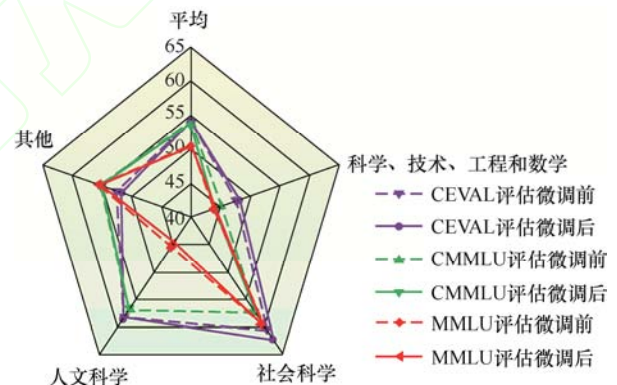


图 6 模型通用效果评估

由模型在 CEVAL 数据集上评估分数可知, 微调后模型的通用中文知识平均性能变化不大, CMMLU 数据集上评分表明模型在科学、技术、工程和数学领域能力有所下降, 但模型整体性能同样变化不大, 而从 MMLU 数据集的评估结果可知, 相对于中文的通用科学、技术、工程和数学知识, 模型的英文通用知识在微调前后相较于中文知识略有不足, 但整体微调并没有明显降低模型的通用知识与推理能力, 模型仍具有开放域知识。

为直观感受领域模型的任务能力, 选择基于特征的工艺问答样例将该领域模型与微调前的模型及 ChatGPT3.5 进行效果比较, 模型表现如表 7 所示。

对于表中的工艺问题，由于领域模型经高质量数据集训练，对于问题回答更精准、专业，ChatGLM3-6B 则无法生成专业的回答，ChatGPT3.5

回答更精细，但信息冗余重点不突出。综上各种结果分析表明，该领域模型可以针对具体零件特征推荐出相应的加工工艺并拥有一定的开放域知识。

表 7 模型问答效果表现

模型	回答
领域模型	问题：根据机械加工的实际情况，给定加工特征与加工参数生成相应加工方法与工艺路线。加工特征：四边形腔。加工参数：140 mm、宽度 120 mm、直径 NULL、深度 62 mm、精度 IT8。 加工方法：粗铣-半精铣-精铣。加工工艺路线：选择 150 mm 长、130 mm 宽、75 mm 深度的铝合金工件材料。在 CNC 加工中心上进行定位和夹紧。使用直径 12 mm 的铣刀进行粗铣，对四边形腔进行粗加工，使其初步成形。更换铣刀，选择直径 6 mm 的铣刀进行半精铣，对四边形腔进行半精加工。最后，更换直径 4 mm 的铣刀进行精铣，以达到加工精度的要求。
ChatGLM3-6B	确定加工工艺路线：由于四边形腔的形状比较复杂，加工难度相对较大，因此我们建议采用分步加工的工艺路线，先进行粗加工，再进行精加工。刀具选择：由于直径未知，我们需要先根据四边形腔的形状和尺寸，以及所使用的刀具类型，估算出所需的刀具直径。建议选择钻头或铣刀作为刀具。加工步骤：进行粗加工：先使用较大的刀具进行初步的加工，以形成四边形腔的基本形状。这一步需要控制好加工的深度和宽度，以确保四边形腔的尺寸符合要求。进行精加工：使用较小的刀具进行精加工，以达到所需的尺寸和精度。这一步需要进行多次加工，逐步完善四边形腔的形状和表面质量。加工完成后，需要进行检验和验收，确保加工结果符合设计要求。
ChatGPT3.5	根据提供的加工特征和参数，对四边形腔的加工方法与工艺路线进行初步规划： 1. 材料准备：选择适当的金属材料，考虑到加工特征和要求的精度。 2. 铣削加工：使用数控铣床，采用合适的铣刀。首先进行面铣，将工件的底面和侧面平整。接着进行轮廓铣削，根据提供的长度和宽度进行相应切削。 3. 开孔加工：针对四边形腔的角部，采用合适的钻头进行开孔。控制深度，确保达到所需的 62 mm 深度。 4. 粗加工与精加工：进行粗加工以快速去除多余材料。随后进行精加工，控制加工参数以满足 IT8 的精度要求。 5. 检验与调整：进行工件的检验，确保加工尺寸和形状符合要求。如有需要，进行微调以满足精度标准。 6. 表面处理：考虑是否需要表面处理，如抛光或镀层，以提高表面光洁度和耐腐蚀性。 7. 清洁与包装：清理工件，去除加工产生的切屑和污物。进行适当包装，保护工件在运输和存储过程中不受损。

3 工艺知识问答推荐系统的构建方法

由于每个数控加工人员对零件关键特征工艺参数设置存在偏差且工艺决策不一，导致相同零件在加工质量上存在工艺一致性差的问题。为此，使用 LangChain 框架将领域模型与加工工艺知识向量数据库进行融合，构建工艺知识问答推荐系统，实现工艺知识的增强，规范零件关键特征工艺的一致性选择。同时，通过向量数据库的融合，解决模型中缺乏最新专业知识及回答结果不精确的问题。在实际应用过程中，主要注重模板提示词的设计，以充分激发领域模型能力并约束系统输出内容与格式。

3.1 LangChain 与向量数据库的结合

图 7 展示了使用 LangChain 与已建立的加工工艺向量数据库结合应用的流程。

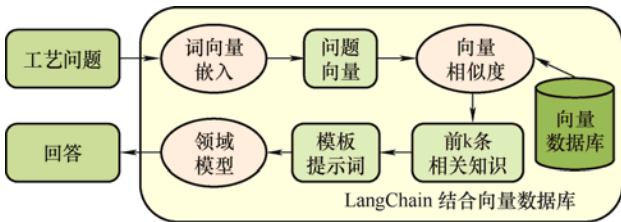


图 7 向量数据库与领域模型的融合应用

1) 对输入的加工工艺问题进行向量化处理，使用智源研究院开源的 bge-large-zh 词向量嵌入模型进行句向量的获取。相较于 Word2Vec、Glove 等传

统静态词向量模型，该模型在中文语义向量综合表征与检索精度方面均拥有不俗的表现。

2) 在获取向量数据库中的语义向量之后，采用 Faiss 库进行相似度搜索。为提高模型回答的效率，在处理向量数据库中的大规模语义向量时，使用 Faiss 库中的倒排暴力检索方法进行向量索引，获取与问题相关的案例及背景知识。

3) 将向量数据库中检索到的前 k 条相关知识与用户输入工艺问题结合，构造适当提示模板。提示模板主要用来生成提示词并引导大语言模型从答案参考中进行理解分析并生成相对准确的答案。

提示词的参考构造如表 8 所示，第一条提示词用于无向量数据库连接情况下，模型根据自身领域能力生成答案；第二条提示词根据向量数据库中获得的知识与工艺问题进行融合提示，激发模型的理解推理能力；第三条提示词在向量数据库中没有搜索到相关知识时候使用，用于约束模型进行胡编乱造并给出相应参考回答。

表 8 模型问答提示词

序号	提示词
1	作为一名资深机械加工工程师回答以下问题，给定下列<加工特征与加工参数>生成相应加工方法与工艺路线
2	作为机械工程领域的专家，简洁和专业地回答问题。根据以下参考知识：“<参考文本>”，结合<参考文本>与<工艺问题>生成相应加工方法与工艺路线
3	未找到相关知识，禁止胡编乱造。接下来由本模型结合自身能力理解问题并为你生成参考解答

3.2 系统构建与评估

通过向量数据库与领域模型的融合构建, 利用 Python 语言与 LangChain 框架搭建起网页端问答系统。该问答系统主要包含知识库管理与对话问答两个模块, 知识库管理整合向量数据库的构建流程用于将工艺知识储存至向量数据库并提供相关数据管理索引功能, 对话模式分为 LLM 对话与知识库对话两种模式, 分别提供领域模型的问答与结合知识库后领域模型的问答。

系统搭建完成后, 为实现系统生成内容的高质量评估, 抛弃之前采用在词粒度上的内容评价指标 BLEU、ROUGE, 采取基于语言模型的评价指标 BERT Score^[28]对测试数据进行测试, 其在句粒度上对仅微调后的领域模型和融合向量数据库的领域模型进行性能对比评估。该指标采用了基于上下文嵌入的相似性计算方法, 通过 BERT 模型将生成文本与参考文本转化为 token(文本中最小的语义单元)序列, 提取特征信息。接着, 计算两段文本中每个 token 的内积, 从而构建一个相似性矩阵。最后, 通过计算候选句子与参考句子中每个 token 标记的相似性, 得出最终的评分, 计算示意图如图 8 所示。

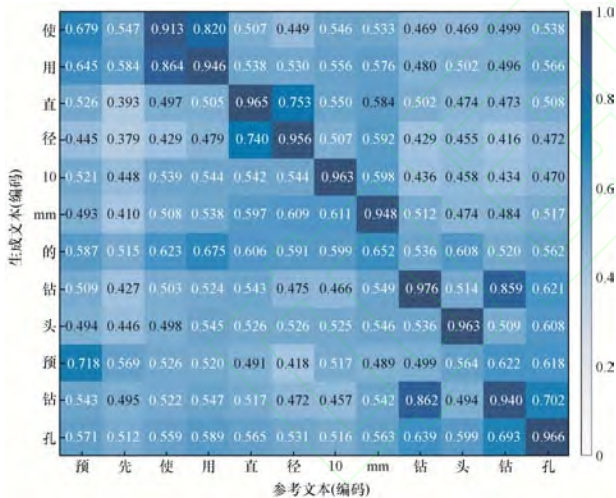


图 8 相似度矩阵示意图

根据评价相似性矩阵得到最终准确率 (Precision)、召回率 (Recall) 和 F1 值如表 9 所示。结果表明, 仅模型微调在盲孔与外圆特征工艺问题问答上整体 F1 值达到 0.9 以上, 表明领域模型可以推荐简单特征工艺知识, 总体 F1 值达到 0.868, 说明模型微调激发了领域模型在该工艺问答任务上的能力, 提升了其任务适配性。向量数据库的引入为领域模型增添了新的工艺知识参考, 使系统问答效果整体提升。但针对具体盲孔、外圆特征的问答结果 F1 值甚至出现下降表现, 可能由于实际背景工艺知

识多样化导致模型理解的偏差与回答的多样性, 但总体 F1 值达到 0.894, 说明向量数据库的引入有一定的优势。

表 9 模型效果评价结果

特征类型	仅模型微调			模型微调+向量数据库		
	Precision	Recall	F1	Precision	Recall	F1
四边	0.837	0.842	0.839	0.880	0.890	0.885
形腔						
方形	0.851	0.864	0.857	0.882	0.855	0.868
槽						
圆形	0.804	0.902	0.850	0.883	0.892	0.887
凸台						
方形	0.866	0.912	0.888	0.947	0.917	0.932
凸台						
通孔	0.791	0.875	0.831	0.896	0.872	0.884
盲孔	0.895	0.925	0.910	0.913	0.885	0.899
外圆	0.920	0.886	0.903	0.890	0.914	0.902
总计	0.852	0.887	0.868	0.899	0.889	0.894

综上分析表明, 微调后的领域模型具备理解基于特征的机械加工工艺问题, 但在答案生成上缺乏工艺知识的多样性与严谨性。通过向量数据库的引入可以为模型获取最新专业知识与工艺案例, 保证推荐结果的准确性。

为进一步评价大语言模型的推荐效果, 衡量大语言模型生成加工工艺的准确度, 本文设计出加工方法匹配度 (Matching degree of processing method, MDPM), 加工刀具匹配度 (Matching degree of machining tools, MDMT) 两个指标对模型生成的单个内容准确度进行测试, 其计算公式见式(1)和式(2)

$$\text{MDPM} = \begin{cases} 1, & m \in V(F, I, R) \\ 0, & m \notin V(F, I, R) \end{cases} \quad (1)$$

$$\text{MDMT} = \begin{cases} 1, & \text{MAX}(d^*) = d \quad (\text{加工方法为钻削时}) \\ 1, & d^* \subseteq D(F, I, R) \quad (\text{加工方法为铣削时}) \\ 0, & \text{其他} \end{cases} \quad (2)$$

式(1)中, m 表示模型生成的加工方法, $V(F, I, R)$ 属于在给定输入特征 F , 加工精度 I 和粗糙度 R 情况下预设的加工方法集合。通过 MDPM 可以评价模型生成单条内容加工方法的准确度。式(2)中, 加工方法为钻削时, d^* 表示模型生成的加工工艺中所用刀具直径集合, d 则表示输入的孔加工特征的直径; 加工方法为铣削时, 则使用 $D(F, I, R)$ 表示在输入加工特征、加工精度与粗糙度条件下所需的加工刀具直径集合。通过 MDMT 可以评价模型生成加工刀具的准确度。通过上述评价指标选取 1 000 条测试数据对不同模型进行对比分析, 平均指标值结果如下表

10 所示。

表 10 模型生成内容平均指标评价表

模型	Avg(MDPM)	Avg(MDMT)
ChatGLM3-6B	94.2	37.4
ChatGPT 3.5	96.4	72.6
仅模型微调	98.6	88.2
模型微调+向量数据库	99.2	90.3

结果表明,各模型在加工方法的生成准确度上均有较高水平,但针对加工刀具的生成,ChatGLM3-6B 与 ChatGPT3.5 的达标率均明显偏低。由于这些模型是由通用知识训练的,因此无法精确针对本领域的特征场景给出准确结果。相对于 ChatGLM3-6B、ChatGPT3.5 与微调后的领域模型,融合向量数据库的问答推荐系统在加工方法与刀具生成的准确度上达标率均在 90% 以上,为后续的实际应用提供了有效的手段。

4 应用案例

4.1 大模型辅助的快速编程系统

在航空结构件数控编程中,应用本文所提的方法搭建了大语言模型辅助的快速编程系统。系统包含基于 LangChain 的加工工艺问答子系统与快速编程模块,加工工艺问答子系统中向量数据库使用实际加工工艺案例及工艺参数进行优化。快速编程系统通过加工工艺问答子系统 API 调用,将其集成到基于 CATIA 的航空结构件快速编程模块中,实现了快速辅助工艺决策。快速编程模块使用 CATIA V5 R21 作为系统开发平台,利用 Visual Studio 2008 开发工具与 CAA 二次开发平台完成开发,主要由特征识别、工艺决策与刀路生成子模块组成,如图 9 所示。

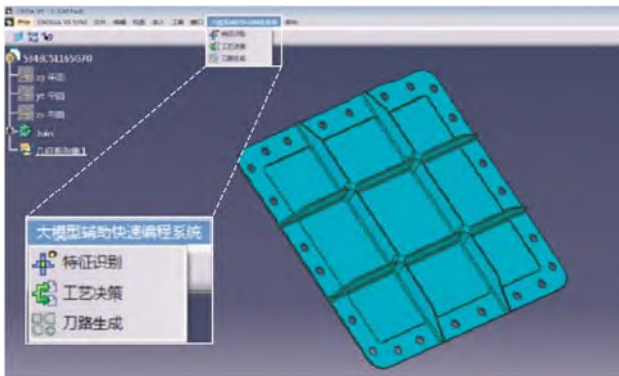


图 9 大模型辅助快速编程系统界面

该系统的特征识别子模块负责将三维模型零件

特征进行提取储存。当特征选择完毕后,传递特征信息交由工艺决策子模块处理,通过结合大语言模型进行针对特征的工艺决策辅助推荐,生成相应加工参数与加工案例并可保存至本地知识库用于后续强化向量数据库进行知识增强。通过加工工艺参数传递至刀路生成子模块,生成特征加工刀轨,最终生成相应的数控加工代码,典型槽特征经该系统处理示例如下图 10 所示。

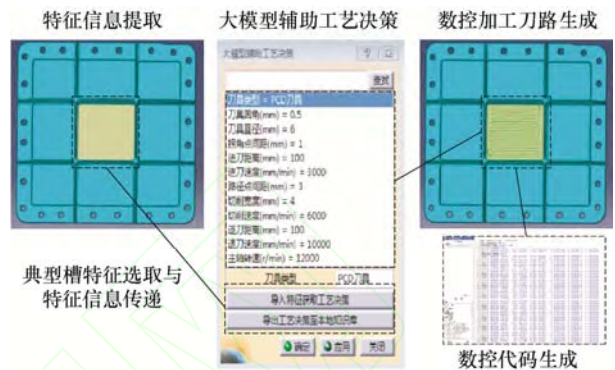


图 10 典型槽特征应用示例

系统中大模型辅助工艺决策模块通过给加工工艺问答子系统提交基于零件特征的规范化模板将工艺问答子系统输出结果规范化,上述示例的规范化模板如下表 11 所示,模板中包含两种类型的提示——详细提示和简短提示。详细提示用于复杂决策场景,提供深入的工艺分析和完整的参数说明;简短提示适用于快速问答场景,提供简洁直接的回答。当接收到加工工艺问答子系统返回的 JSON 格式参数数据后提取相应参数信息至本模块进行参数展示,助力工艺人员进行高效决策。

表 11 模型回答规范化模板

模型提示	提示文本
详细提示	<p>作为机械加工工程师,您将使用知识库案例和已有知识来回答基于零件特征的工艺问题,并输出相应的加工参数。将提供详细的零件特征描述,根据这些特征,您将使用相关工艺知识来确定适当的加工方法和工艺路线。最后,将加工参数以 JSON 格式输出,通过案例参考输出包括主轴转速、进给速度、刀具直径等必要参数。零件特征和要求如下: ``{特征信息}``</p> <p>请按照以下格式输出加工参数:</p> <pre>{ "加工参数 1": "", "加工参数 2": "", "加工参数 3": "", ... }</pre>
简短提示	<p>现在你是{某个领域,如:机械加工}领域的专家。请参考以下数据: {专业知识文本}, 回答问题: {用户输入的问题}</p>

4.2 效果评估

为验证所提方法的应用效果,以典型框体类航空结构件进行数控编程任务为应用测试场景,验证

工艺知识重用及本文所提方法的有效性。在该任务中,针对典型框体类航空结构件进行数控加工编程,选取零件特征复杂度由低到高的三种零件如图 3 所示,完成零件的快速数控编程生成数控编程代码,邀请新入行设计人员使用该大模型辅助的快速编程系统生成数控代码。为进一步分析本文所提方法的实际应用效果,将本文所提系统中大语言模型辅助工艺决策模块去除作为传统自动数控编程方式与该系统进行编程用时对比,结果如表 12 所示。

表 12 大模型辅助数控编程系统数控代码生成效率对比

零件	编程方式	总耗时	效率提升	加工结果
零件 1	本文编程系统 自动数控编程	2 min58 s 4 min32 s	52.81%	合格 合格
零件 2	本文编程系统 自动数控编程	3 min28 s 5 min26 s	56.73%	合格 合格
零件 3	本文编程系统 自动数控编程	4 min20 s 7 min23 s	70.38%	不合格 合格

通过效率对比分析可得,本文基于大语言模型辅助的快速编程系统相比没有使用大语言模型辅助的工艺决策模块进行数控编程,效率提升明显。具体而言,未使用大语言模型辅助的工艺决策模块时,需由人工在 CATIA 中针对不同加工特征选择相应的工艺信息,包括刀具及加工参数(如加工宽度、深度、进给速度)等。特别是对于包含较多复杂特征的零件,所需重复选择较多,导致编程时间显著增加。使用本文所提的大模型辅助数控编程系统,利用向量数据库中的历史知识与大语言模型自身的推理与推荐能力,可以自动化地为不同特征推荐相应的工艺参数,显著减少人工干预,提高编程效率。针对复杂特征较多的零件 3,本文所提系统在生成的数控代码有误,原因在于针对少数复杂特征,大语言模型由于缺少历史经验及模型回答的不稳定性导致生成的工艺参数有误,后续考虑基于因果理论构建数据之间的因果关系图,以此进一步微调训练增强大语言模型的推理能力与回答稳定性。

5 结论

(1) 基于 ChatGLM3-6B 构建机械加工工艺知识推荐领域模型,通过 LoRA 微调技术与其他微调技术的对比选优,得到领域适配微调方法与模型,根据评价指标分析表明,所构建的领域模型能够根据具体零件特征推荐出相应的加工工艺。

(2) 提出领域模型与向量数据库结合的工艺知

识增强方法,通过 BERT Score 指标评估,该方法总体 F1 值达到 0.894,在领域模型基础上提升 0.026。通过 MDPM 与 MDMT 指标评价,该方法的加工工艺推荐准确度在 90% 以上。

(3) 通过航空结构件数控编程应用验证了大语言模型辅助的快速编程系统的可行性,对比未使用本文所提方法的自动数控编程系统,框体类结构件数控代码生成效率在时间上有所提升。

现实航空结构件通常包含较多复杂特征及不同的加工工艺,所需加工工艺推荐具有较高的复杂性。因此,未来研究可以进一步优化大语言模型的微调方法,使用因果理论增强大语言模型推理的可解释性与回答稳定性,并在实际加工环境中进行广泛验证以检验其稳定性与普适性。

参 考 文 献

- [1] 尹佳,唐宇阳,张俊,等. 基于复合加工特征的航空结构件频响快速预测[J]. 机械工程学报, 2023, 59(3): 200-207.
YIN Jia, TANG Yuyang, ZHANG Jun, et al. Rapid frequency response function prediction of aeronautical structural parts based on composite machining features[J]. Journal of Mechanical Engineering, 2023, 59(3): 200-207.
- [2] 常智勇,陶礼尊,李佳佳,等. 基于加工意图的机加工工艺知识重用方法研究[J]. 机械工程学报, 2018, 54 (3): 160-168.
CHANG Zhiyong, TAO Lizun, LI Jiajia, et al. The measure and search method of process knowledge element based on machining intent[J]. Journal of Mechanical Engineering, 2018, 54(3): 160-168.
- [3] ZHOU D, DAI X. A method for discovering typical process sequence using granular computing and similarity algorithm based on part features[J]. The International Journal of Advanced Manufacturing Technology, 2015, 78(9): 1781-1793.
- [4] 罗滨鸿,周虎,张祺薇,等. 基于 Apriori 算法的网线编织工艺缺陷数据挖掘方法[J]. 制造业自动化, 2022, 44(5): 75-77+102.
LUO Binhong, ZHOU Hu, ZHANG Qiwei, et al. A data mining method for defect data of net rope weaving process based on apriori algorithm[J]. Manufacturing Automation, 2022, 44(5): 75-77+102.
- [5] 阳树梅,王华昌,李建军. 基于经验的数控工艺知识挖掘算法研究[J]. 模具工业, 2023, 49(8): 1-10.
YANG Shumei, WANG Huachang, LI Jianjun. Research on knowledge mining algorithm of NC process based on

- experience[J]. Die & Mould Industry, 2023, 49(8): 1-10.
- [6] ZHOU B, BAO J S, LI J, et al. A novel knowledge graph-based optimization approach for resource allocation in discrete manufacturing workshops[J]. Robotics and Computer-Integrated Manufacturing, 2021, 71(3): 102160-102173.
- [7] ZHENG P, XIA L Q, LI C X, et al. Towards Self-X cognitive manufacturing network: an industrial knowledge graph-based multi-agent reinforcement learning approach[J]. Journal of Manufacturing Systems, 2021, 61(1): 16-26.
- [8] ZHOU B, LI X, LIU T, et al. CausalKGPT: industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing[J]. Advanced Engineering Informatics, 2024, 59: 102333.
- [9] 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.
- [10] DING N, QIN Y J, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.
- [11] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners[J]. Advances in Neural Information Processing Systems, 2022, 35: 22199-22213.
- [12] LIU Xiao, JI Kaixuan, FU Yicheng, et al. P-Tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks[M/OL]. arXiv, 2021[2023-12-6]. <https://arxiv.org/abs/2110.07602>.
- [13] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models[M/OL]. arXiv, 2021[2023-12-1]. <https://arxiv.org/abs/2106.09685>.
- [14] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: efficient finetuning of quantized llms[M/OL]. arXiv, 2023[2023-12-8]. <https://arxiv.org/abs/2305.14314>.
- [15] 张鹤译, 王鑫, 韩立帆, 等. 大语言模型融合知识图谱的问答系统研究[J]. 计算机科学与探索, 2023, 17(10): 2377-2388.
- ZHANG Heyi, WANG Xin, HAN Lifan, et al. Research on question answering system on joint of knowledge graph and large language models[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(10): 2377-2388.
- [16] 赵鑫, 窦志成, 文继荣. 大语言模型时代下的信息检索研究发展趋势[J]. 中国科学基金, 2023, 37(5): 786-792.
- ZHAO Xin, DOU Zhicheng, WEN Jirong. The development of information retrieval in the era of large language model[J]. Bulletin of National Natural Science Foundation of China, 2023, 37(5): 786-792.
- [17] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172-180.
- [18] HUAGN A H, WANG H, YANG Y. FinBERT: A large language model for extracting information from financial text[J]. Contemporary Accounting Research, 2023, 40(2): 806-841.
- [19] 张俊, 徐箭, 许沛东, 等. 人工智能大模型在电力系统运行控制中的应用综述及展望[J]. 武汉大学学报(工学版), 2023, 56(11): 1368-1379.
- ZHANG Jun, XU Jian, XU Peidong, et al. Overview and prospect of application of artificial intelligence large model in power system operation control[J]. Engineering Journal of Wuhan University, 2023, 56(11): 1368-1379.
- [20] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: open and efficient foundation language models[M/OL]. arXiv, 2023[2023-12-20]. <https://arxiv.org/abs/2302.13971>.
- [21] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
- [22] TAORI R, GULRAJANI I, ZHANG Tianyi, et al. Stanford alpaca: an instruction-following llama model[EB/OL]. GitHub Repository, 2023[2023-12-12]. https://github.com/tatsu-lab/stanford_alpaca.
- [23] DU Zhengxiao, QIAN Yujie, LIU Xiao, et al. General language model pretraining with autoregressive blank infilling[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022: 320-335.
- [24] GUU K, LEE K, TUNG Z, et al. Retrieval augmented language model pre-training[C]// Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020: 3929-3938.
- [25] PAPINENI K, ROUKO S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: ACL, 2002: 311-318.
- [26] LIN C Y. Rouge: a package for automatic evaluation of summaries[C]// Text summarization branches out.

Barcelona, Spain: ACL, 2004: 74-81.

- [27] KORBAC T, ELSAHAR H, KRUSZEWSKI G, et al. Controlling Conditional Language Models without Catastrophic Forgetting[C]// Proceedings of the 39th International Conference on Machine Learning. Baltimore, Maryland, USA: PMLR, 2022, 162: 11499-11528.
- [28] ZHANG Tianyi, KISHORE V, WU F, et al. Bertscore:

evaluating text generation with bert[M/OL]. arXiv, 2019[2023-12-26]. <https://arxiv.org/abs/1904.09675>.

作者简介: 郑小虎(通信作者), 男, 1983 年出生, 博士, 副教授, 硕士研究生导师。主要研究方向为设备智能运维技术、知识图谱、深度学习、智能工艺与仿真。

E-mail: xhzheng@dhu.edu.cn

陈宏博, 男, 2000 年出生, 硕士研究生。主要研究方向为数字孪生和大语言模型。

E-mail: hbchen@mail.dhu.edu.cn