ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 637－646
DOI:10. 16337/j. 1004-9037. 2025. 03. 006

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

武信:一种垂直领域大语言模型系统架构设计与实证

朱新立[1,2], 高志强[2], 姬纬通[2], 李少华[2], 李松杰[2]
(1.北方自动控制技术研究所,太原030006;2.武警工程大学重点实验室,西安 710086)

摘　要: 在定制化应用场景下亟需提升大语言模型(Large language models, LLMs)在特定垂直领域的语言理解和生成能力。本文提出一种适用于垂直领域的大语言模型系统开发范式——武信。其涵盖架构、数据、模型和训练等大语言模型系统的系列开发方法,利用人在回路的数据增强提升军事训练伤问答数据集的质量,采用梯度低秩投影(GaLore)策略对轻量级基座大语言模型进行高效全参微调。实验结果表明,所采用的全参微调方法在收敛性和准确性指标上优于主流的LoRA微调,所训练的武信大模型在军事训练伤防治专业知识理解、克服"幻觉"等方面优势明显,相关成果可为垂直领域问答大模型系统设计与应用提供参考。

关键词: 数据增强;大语言模型系统;全参微调;垂直领域大模型
中图分类号: TP391.1　　**文献标志码:** A

Wuxin: Architecture Design and Empirical Study for Vertical-Domain Large Language Model System

ZHU Xinli[1,2], GAO Zhiqiang[2], JI Weitong[2], LI Shaohua[2], LI Songjie[2]
(1. North Automatic Control Technology Institute, Taiyuan 030006, China; 2. Key Laboratory of CTC &IE (Engineering University of PAP), Ministry of Education, Xi'an 710086, China)

Abstract: In customized scenarios, it is urgent to enhance the understanding and generation capabilities of large language models (LLMs) in specific vertical domains. We propose a paradigm for developing vertical-domain LLM system named "Wuxin", which covers a series of development methods for LLM systems, including architecture, data, model, and training. Wuxin utilizes human-in-the-loop data augmentation to improve the quality of military training injury question and answer datasets, and employs the GaLore strategy to perform efficient full-parameter fine-tuning on small LLMs. Experimental results show that the adopted full-parameter fine-tuning method outperforms LoRA fine-tuning in terms of convergence and accuracy. Furthermore, Wuxin demonstrates significant advantages in understanding professional military training injury knowledge, as well as overcoming hallucinations. Our achievements can provide references for the design and application of question-answering LLM systems in vertical domains.

Key words: data augmentation; large language model system; full-parameter fine-tuning; vertical-domain large model

---

基金项目:国家社会科学基金(2022-SKJJ-C-093);武警部队科技创新团队创新研究项目(ZZKY20222103)。
收稿日期:2025-01-26;**修订日期:**2025-03-21

引　言

大语言模型等颠覆性人工智能技术正以前所未有的广度与深度影响着新一轮智能技术的变革。同时,大语言模型以其强大的语言理解和生成能力,在自然语言处理领域取得了显著的进步。2017年,谷歌推出用于处理自然语言任务的Transformer网络架构,次年,OpenAI发布了GPT-1(Generative pre-trained Transformer-1),能够生成流畅的自然语言文本[1]。2022年,OpenAI发布的大语言模型ChatGPT[2]具有极高的人机交互水平和极为广泛的应用场景,引起了全社会的广泛关注。这些大模型如BERT(Bidirectional encoder representations from Transformers)、GPT、T5等[3],通常基于Transformer架构,并在大规模无标签文本数据上进行预训练,学习通用的语言表示,进而通过监督学习在下游任务上进行微调,取得了优异的自然语言理解与生成效果。

在特定的垂直应用领域,基于大语言模型的人工智能问答系统在训练伤防治领域具有广泛的应用需求。例如,体育训练是增强体能素质的基本途径,然而在训练实际过程中,组织方法不科学、防护措施不到位、训练动作不规范以及心理素质欠佳等因素导致的训练伤频发,尤其军事训练高强度、高对抗和高要求的特殊封闭场景,训练伤防治更是与战斗力生成密切相关。因此,针对军事训练伤防治等智能问答需求,研发大模型高效微调技术及系统架构,降低计算资源消耗、提升系统回答精度,激发大模型在垂直领域问答的潜能,保障官兵军事训练安全,在军事训练伤防治领域实现更专业、更有效的问答"助手"系统,具有重要的现实意义和推广价值。

本文首先梳理了垂直领域大模型的研究现状、代表性大模型及实际应用;其次介绍了武信大语言模型系统的整体框架、信息流程以及人在回路的数据增强、GaLore策略全参数微调等系列关键技术方法;最后从收敛性与准确性维度评估武信大模型的全参数微调性能,并完成军事训练防治问答方面案例实证的评估分析。

1　垂直领域大模型研究现状

从早期的统计语言模型(Statistical language model, SLM)和基于神经网络的语言模型(Neural language models, NLM)的初步探索,到当前基于Transformer架构的预训练语言模型(Pre-trained language model, PLM)以及大语言模型的深入研究(如图1所示),人工智能技术在自然语言处理领域取得了显著进展。OpenAI发布的GPT-1在传统语言模型的预训练基础上进行了微调,GPT-3引入上下文学习机制的提示式微调方法,InstructGPT采用了基于指令的微调方法,GPT-4基于人类反馈的强化学习技术,并将输入由单一文本模态扩展到了图文双模态。Meta开源的Llama-3.1-405B在常识、数学、工

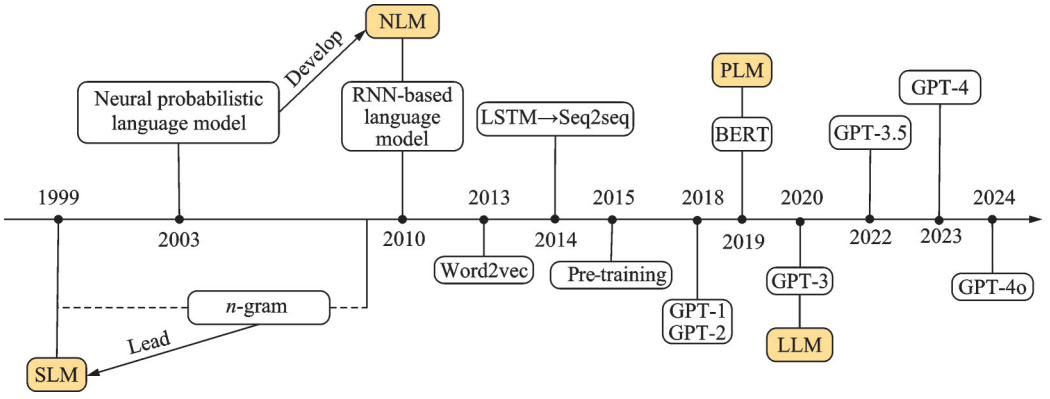

图1　语言模型发展历程

Fig.1　Development process of language models

具使用和多语言翻译等系列任务中,可与GPT-4o、Claude-3.5-Sonnet和Gemini-Ultra相媲美。国内“文心一言”大模型、抖音云雀大模型、阿里通义千问大模型、月之暗面Kimi、科大讯飞星火大模型等如雨后春笋般蓬勃发展[4-6]。

除了通用人工智能大语言模型外,垂直领域大模型作为人工智能技术中的一种重要形式,逐渐成为了医疗、金融、教育和交通等行业的焦点。微脉的“全病程管理”、深睿医疗的“Deepwise MetAI”以及深圳市大数据研究院的“华佗GPT”等医疗领域大模型可以帮助医生进行疾病诊断、治疗方案制定,提高医疗服务的准确性和效率;度小满推出国内首个金融行业的开源大模型“轩辕”,蚂蚁集团发布自研的“贞仪”语言和多模态大模型,恒生电子发布了金融行业大模型“LightGPT”;科大讯飞的“星火语伴”、好未来的“Math GPT”以及可汗学院的“Khanmigo”等教育领域大模型可以用于个性化教学、智能答疑等,提高教学效果和学习体验。尤其,在医疗领域,大模型有望实现快速准确识别潜在损伤、推荐合适的治疗策略、预测治疗结果,推动实现个性化智能医疗服务[7]。然而,如何利用特定场景的本地化私有数据,结合通用基座大语言模型的基础理解和生成能力,实现定制化应用场景下的特定垂直领域快速全参微调、高效训练,已成为大语言模型系统提升特定应用领域语言理解和生成能力的重要研究方向。

2 武信:大语言模型系统架构设计

2.1 整体架构

本文提出一种适用于定制化应用场景下的大语言模型系统开发范式,即“武信”大语言模型系统架构,其核心为“数据+模型+训练+问答”模式,聚焦军事训练伤防治领域应用需求,进行定制化军事训练伤防治的数据贯通、全参微调和问答交互,形成了完整的数据基座、数据增强、全参微调和垂直领域应用的通用大模型系统开发范式与定制化应用的全栈链路。如图2所示,武信大语言模型系统架构具有多层次、模块化和垂直领域微调训练、智能问答等特点,自顶向下由灵活可扩展的问答层、训练层、模型层和数据层组成。其中,B是通用的大模型参数量,B为billion,代表十亿。

图2 整体架构

Fig.2 Overall architecture

2.1.1 数据层

数据层聚焦训练伤防治数据的多样性和覆盖面,包括严格筛选和预处理的开源中文医疗数据集cMedQA2[8]和人工数据增强的领域数据。cMedQA2数据集是华人社区医学问答数据集2.0版本,覆盖了广泛的医疗知识问答,分为训练数据、验证数据和测试数据,共计问题108 000条、答案203 569条,如表1所示。根据《国际疾病分类

表1 cMedQA2数据集

Table 1 cMedQA2 dataset

数据集	问题数	答案数	平均词数(问题)	平均词数(答案)
训练集	100 000	188 490	48	101
验证集	4 000	7 527	49	101
测试集	4 000	7 552	49	100

第十一版(ICD-11)》和《军事训练医学指导手册》标准框架,以及医学专家主导的三级校验机制,经过格式标准化、异常值修正等规范化操作后,从cMedQA2数据集的原始数据中严格筛选出2 000余条高质量问题-答案对,同时通过多轮人工校验和验证,形成了区分骨骼肌肉损伤、皮肤软组织伤和环境因素伤等8大损伤类别,具体包含骨折、膝关节半月板损伤和腰椎间盘突出等42种常见训练伤病类型,以及特殊军事训练相关的跟腱断裂、跟腱炎等损伤类型的训练数据基座。

在针对军事训练伤防治数据的定制化增强方面,整合互联网公开的相关科研文献,例如,《解放军医学杂志》《军事体育研究》等期刊成果,再次利用人在回路的数据标注、专家审核等优选方法筛选,在cMedQA2数据集的分类体系基础上,增强扩充出500余条结构化的问题-答案对,重点强化肌肉拉伤、关节损伤和军事训练伤防治三类应用场景数据,扩大训练数据在急性损伤处置、慢性劳损康复等关键应用场景的覆盖率,确保军事训练伤防治数据集的质量和分析结果的可靠性。部分高质量人在回路增强的问题-答案对如图3所示。

图3 部分高质量问题-答案对

Fig.3 Partial high-quality question-answer pairs

2.1.2 模型层

模型层的设计充分考虑了模型参数量的差异性和计算资源需求的互补性,综合不同用户的算力约束和应用场景需求,基于Qwen2在语言理解、语言生成、编码和推理等一系列能力[9],武信大模型系统的基础模型支持Qwen2-0.5B、Qwen2-1.5B、Qwen2-7B、Qwen2-72B等,在训练微调后,进一步提升所适配的预训练模型在特定任务上的性能。其中,Qwen2-0.5B-Instruct是Qwen2系列中具有0.5亿参数的指令微调型语言模型,在Transformer架构基础上,通过SwiGLU激活函数来增强模型的表达能力,在32k上下文长度上进行训练,可通过Dual chunk attention等技术扩展至更长的上下文长度。此外,Qwen2-0.5B-Instruct采用Query-Key-Value(QKV)偏置和组查询注意力提升模型更好地捕捉长距离依赖关系,提高对复杂语言结构的理解、高效准确的自然语言处理能力,并具有较好的指令遵循能力。

Qwen2-0.5B-Instruct在多个基准测试中展现了语言理解、生成和多语言处理等方面的良好性能。如表2所示,多语言理解(Massive multitask language understanding, MMLU)基准测试的得分为37.9,相较于Qwen1.5-0.5B-Chat的35.0分有显著提升;在HumanEval测试中,Qwen2-0.5B-Instruct的得分为17.1,远高于Qwen1.5-0.5B-Chat的9.1;GSM8K测试的得分达到了40.1,C-Eval测试的得分为45.2,显示了生成任务能力和中文语言理解上的优势;在IFEval的Prompt Strict-Accuracy测试中,Qwen2-0.5B-Instruct的得分为20.0,显示出其在遵循严格指令方面的能力。

表2 Qwen2-0.5B-Instruct数据集评测

Table 2 Evaluation results of Qwen2-0.5B-Instruct datasets

数据集	Qwen1.5-0.5B-Chat	Qwen2-0.5B-Instruct	Qwen1.5-1.8B-Chat
MMLU	35.0	**37.9**	43.7
HumanEval	9.1	**17.1**	25.0
GSM8K	11.3	**40.1**	35.3
C-Eval	37.2	**45.2**	55.3
IFEval	14.6	**20.0**	16.8

2.1.3 训练层

武信大模型系统架构的训练层是大模型性能优化的核心,提供模型参数优化的策略和工具,适配GaLore策略[10]全参微调、LoRA[11]等多种主流微调方法,使大模型可以在低计算资源环境下适应军事训练伤防治等特定应用场景的知识和数据特性。全参微调适用于下游任务与预训练模型差异大或需高度灵活适应的场景,但训练时间长、计算资源消耗大;部分微调适用于目标任务与预训练模型有一定相似性或任务数据集较小的情况,计算资源和时间需求少,但性能可能略有下降,“幻觉”问题突出,主要包括LoRA微调、适配微调、前缀微调、提示微调和P-Tuning等方法[12-13]。在支持LoRA的基础上,武信系统架构的训练层主要采用梯度低秩投影(GaLore)全量参数学习策略,以实现在不牺牲模型性能的前提下显著降低内存消耗。

此外,关于大语言模型的能力评估,主要包括指标自动评估和人工评估两类,前者简单快速,不需要人工参与,节省时间和成本,适用于如数学问题之类的大多数确定性任务;后者适用于开放生成式任务,可以更接近实际应用场景,提供更准确和全面的反馈。具体而言,收敛性和准确性较为常用[14]。其中,准确性作为衡量模型输出和预期输出的匹配程度的指标,可以衡量模型预测或生成结果的正确比例。在相关衍生指标中,准确率是模型对所有数据集进行实验的准确率的平均数。例如,对于生成任务主要有双语评估(Bilingual evaluation understudy, BLEU)、ROUGE(Recall-oriented understudy for gisting evaluation)等指标[15]。

BLEU指标用于衡量准确率。给定标准为reference,模型生成为candidate,长度为n,candidate中有m个单词出现在reference,m/n为BLEU的1-gram的计算公式。根据连续词模型n-gram,BLEU可以划分成多种评价指标,例如,BLEU-1、BLEU-2、BLEU-3、BLEU-4,其公式为

$$\mathrm{BLEU}=\mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N}\omega_n\log P_n\right)\quad \omega_n=\frac{1}{N} \tag{1}$$

$$\mathrm{BP}=\begin{cases}1 & c>r\\ \mathrm{e}^{1-\frac{r}{c}} & c\leqslant r\end{cases} \tag{2}$$

式中:BP为惩罚因子,惩罚输出句子过短,防止训练结果倾向短句;c为候选句子的词数;r为参考句子的词数;ω_n为权重;通常N取1~4。

P_n为n-gram的精确度,其公式为

$$p_n=\frac{\sum_{n\text{-gram}}\mathrm{CounterClip}(n\text{-gram})}{\sum_{n\text{-gram}}\mathrm{Counter}(n\text{-gram})} \tag{3}$$

ROUGE是评估自动文摘以及机器翻译的指标,通过将生成与参考(通常是人工生成的)进行比较计算,得出相应的分值,以衡量自动生成与参考之间的“相似度”。ROUGE-N主要统计n-gram上的召回率,对于n-gram,可以计算得到ROUGE-N分数,计算公式为

$$\text{Rouge-}N=\frac{\sum_{S}\sum_{n\text{-gram}}\text{Count}_{\text{math}}(n\text{-gram})}{\sum_{S}\sum_{n\text{-gram}}\text{Count}(n\text{-gram})} \tag{4}$$

ROUGE-L 中 L 指最长公共子序列(Longest common subsequence, LCS),用于计算生成 C 和参考 S 的最长公共子序列,计算公式为

$$\text{Rouge-}L=\frac{(1+\beta^2)R_{\text{LCS}}P_{\text{LCS}}}{R_{\text{LCS}}+\beta^2P_{\text{LCS}}} \tag{5}$$

式中:R_{LCS} 表示召回率;P_{LCS} 表示精确率;β 一般为较大的数。

$$R_{\text{LCS}}=\frac{\text{LCS}(C,S)}{\text{len}(S)}\qquad P_{\text{LCS}}=\frac{\text{LCS}(C,S)}{\text{len}(C)} \tag{6}$$

式中:len(•)表示求字符中字符数的长度函数;S、C 表示字符中字符数。

2.1.4 问答层

问答层是武信大语言模型系统与用户直接交互的接口,负责接收用户的自然语言输入并提供相应的输出,例如,用户关于军事训练伤防治领域的问题输入和大语言模型的答案生成输出。在武信系统架构的信息处理流程中,全参微调训练后的大语言模型支持对特定领域知识库与通用领域数据的深度融合如图4所示,系统接收用户输入的自然语言问题,进入大语言模型基于Transformer的编码器-解码器架构,完成从输入问题到高维词向量的编码转换。编码后的词向量能够捕捉问题的语义信息,大语言模型利用预测机制和上下文信息,生成预测输出。此外,基于领域特定知识库以及通用领域等知识库,可为问答过程中提供辅助信息,增强模型对特定领域术语和概念的理解。

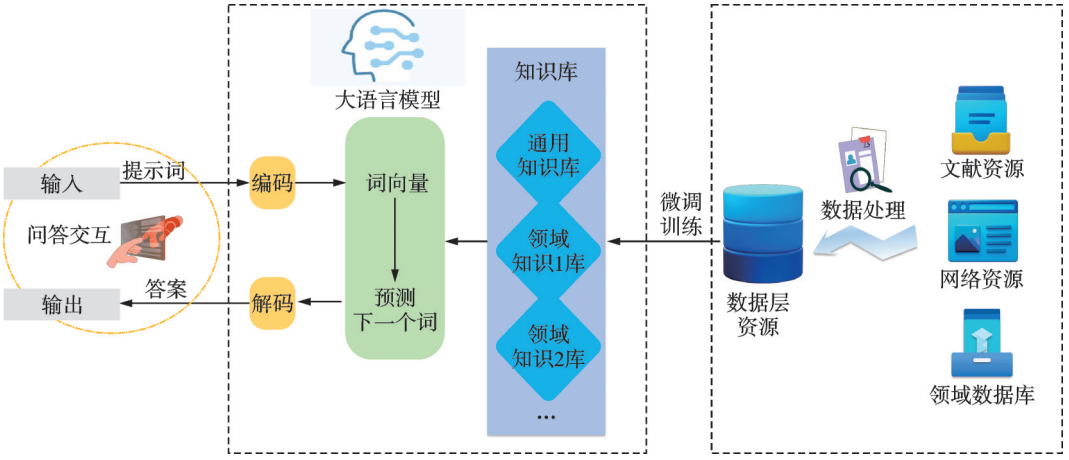

图4 信息处理流程

Fig.4 Information processing flow

2.2 基于GaLore策略的全参微调训练

目前,LoRA是参数高效微调最常用的方法,但LoRA的超参数需要根据模型复杂性、适应需求及欠拟合或过拟合风险进行调优,以找到适合新任务的计算需求与训练效率的最优平衡。经过多种微调方法实践,武信系统架构采用基于GaLore策略的全参微调方法训练,形成了影响模型准确性和收敛速度的系列经验超参数,例如训练轮数、批量大小、学习率和Rank等。相比于LoRA方法将全参微调的增量参数矩阵 $\Delta\boldsymbol{W}$ 表示为两个参数量更小的矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的低秩近似,GaLore策略本质是利用权重矩阵 $\boldsymbol{W}$ 的梯度 $\boldsymbol{G}(\boldsymbol{W}\in\mathrm{R}^{m\times n})$ 缓慢变化的低秩结构,而不是将权重矩阵本身近似为低秩。

GaLore策略的数学理论基础为,梯度矩阵G的秩在训练过程中会变低,即梯度矩阵可以通过较小的子空间近似表示,通过计算两个投影矩阵$P\in\mathrm{R}^{m\times r}$和$Q\in\mathrm{R}^{n\times r}$,将梯度矩阵$G$投影到一个低秩形式$P^{\mathrm{T}}GQ$,有

$$W_T=W_0+\eta\sum_{t=0}^{T-1}\widetilde{G}_t\qquad \widetilde{G}_t=P_t\rho_t(P_t^{\mathrm{T}}G_tQ_t)Q_t^{\mathrm{T}} \tag{7}$$

式中:W_0为预训练权重矩阵;P_t为逐层权重更新参数;η为学习率。

由于在复杂任务上,单个低秩子空间难以捕获整个梯度轨迹。因此,如图5所示,GaLore策略通过建立多个子空间并在训练期间切换不同的子空间来进行全参学习。

$$W_t=W_0+\Delta W_{T1}+\Delta W_{T2}+\Delta W_{Tn} \tag{8}$$

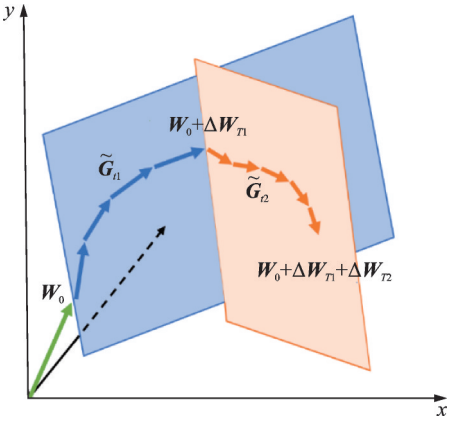

图5 GaLore的低秩子空间学习

Fig.5 Low-rank subspace learning of GaLore

因此,在全参训练大语言模型时,GaLore策略仅需计算两个投影矩阵P和Q,即可以大幅降低依赖于梯度统计量的优化器状态的内存成本。相比于LoRA方法,GaLore策略更适用于专业术语密集的垂直领域优化,为军事训练伤防治任务微调策略选择提供了理论依据。

3 实验分析

3.1 实验环境

实验案例环境配置单卡NVIDIA A10-24 GB GPU,所需软件依赖库及实验推荐版本如表3所示。大模型微调训练工具采用LLaMA-Factory平台,训练数据采用OpenAI格式,如图6所示。

表3 依赖项及版本号

Table 3 Environment dependencies and version number

名称	版本号	名称	版本号
Python	3.10.14	Accelerate	0.32.1
Torch	2.3.0	Peft	0.11.1
Transformers	4.42.4	Trl	0.9.6
Datasets	2.18.0	Tensorflow	2.17.0

```
[
  {
    "messages": [
      {
        "role": "system",
        "content": "系统提示词（选填）"
      },
      {
        "role": "user",
        "content": "人类指令"
      },
      {
        "role": "assistant",
        "content": "模型回答"
      }
    ]
  }
]
```

图6 训练数据格式

Fig.6 Training data format

3.2 效果评估

3.2.1 收敛情况分析

以Qwen2-0.5B-Instruct为基础模型,比对GaLore策略全参微调和LoRA微调两种方式。微调超参数的经验为:较低的学习率有助于模型更好地学习新任务,同时保留之前任务的知识,较大的批量大小会导致模型更容易遗忘之前任务的知识。因此,通过调节学习率、批量大小和训练轮数,调整训练过程中训练参数量、训练时间和显存占用量,优化模型性能,实验中超参数设置如表4所示。

通过设置不同的Rank值,得到的GaLore策略全参微调和LoRA微调的训练时间、显存占用和损失对比结果如表5所示。从训练时间来看,随着GaLore策略的Rank值增加,训练时间略有增加,但总体变化不大,保持在11 min左右。这表明GaLore策略在不同Rank值下的训练效率相对稳定。而LoRA的训练时间明显更长,达到了17.66 min,这可能是因为LoRA在微调过程中需要更多的计算资源。

显存占用方面，随着GaLore策略的Rank值增加，显存占用也相应增加，这是因为更高的Rank值意味着模型参数的增加，需要更多的显存来存储这些参数。LoRA在相同Rank值下显存占用略高于GaLore，这可能是因为LoRA模型的参数结构或者优化算法导致显存使用效率略有不同。

此外，损失是衡量模型性能的一个重要指标，损失越低，通常意味着模型的预测准确性越高。从数据来看，GaLore策略随着Rank值的增加，损失逐渐降低，这表明模型的预测准确性随着参数数量的增加而提高。而LoRA在相同Rank值下损失较高，这可能意味着LoRA模型在当前的微调策略下没有达到最佳的性能。

综合来看，GaLore策略在不同Rank值下表现出较好的训练效率和显存使用效率，且随着Rank值的增加，模型的预测准确性提高。LoRA在相同Rank值下训练时间较长，显存占用略高，且损失较高，表明LoRA在当前的微调策略下性能不如GaLore策略。

表4 训练参数表

Table 4 Training parameters

超参数	值	超参数	值
学习率	1e−5	精度	bf16
训练轮数	10	最大长度	512
批量大小	8	学习率调度器	Cosine

表5 训练效果分析

Table 5 Analysis of training effects

训练方法	训练时间/min	显存占用/GB	损失
GaLore(Rank=128)	11.09	17.45	1.949
GaLore(Rank=512)	11.39	19.42	1.648
GaLore(Rank=1 024)	11.72	20.65	1.504
LoRA (Rank=1 024)	17.66	21.33	2.832

图7详细展示了整个训练过程的损失曲线。X轴表示训练过程中的迭代次数，Y轴表示在每个全局步数下模型的损失值。可以看出，GaLore策略训练损失下降速度更快，且4种不同的微调模型在训练120步时均趋于平稳，表明在当前训练条件下GaLore策略收敛速度更快并与LoRA同时达到收敛状态；同时，训练损失的下降趋势和最终值是评估模型性能的重要指标，GaLore策略在任意全局步数下的损失值都远低于LoRA，这表明即使使用低秩的GaLore策略，其效果也远超于高秩的LoRA。

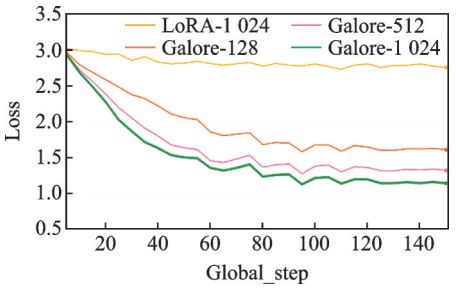

图7 训练损失对比

Fig.7 Comparison of training losses

3.2.2 准确性分析

表6为GaLore策略全参微调与LoRA微调在评估集上的BLEU和ROUGE数据指标。针对BLEU和ROUGE的评估结果显示，GaLore策略随着Rank值的增加，在所有评估指标上都有显著提升，尤其是BLEU-4和ROUGE-2，这表明在处理复杂文本结构和短语生成方面的能力增强。此外，实验中发现，GaLore策略的Rank为1 024时，评测指标却低于Rank为512，尽管Rank值的增加会增加参数数量，理论上应该能够捕捉更复杂的模式，但Rank为512时，指标数据表现最佳，这可能是由于以下3个因素。(1)过拟合。增加模型的参数数量可以提高模型的表达能力，但同时也增加了过拟合的风险。当模型过于复杂时，可能会过度适应训练数据中的噪声和细节，而不是学习到泛化的模式，导致在测试集或实际应用中的表现下降。(2)训练不充分。对于更复杂的模型(如Rank=1 024的模型)，可能需要更多的训练数据和更长的训练时间

表6 不同模型BLEU和ROUGE指标比对

Table 6 BLEU and ROUGE comparison of different models

模型	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
GaLore(Rank=128)	12.98	28.64	13.21	24.63
GaLore(Rank=512)	30.58	42.25	30.33	39.61
GaLore(Rank=1 024)	29.59	41.57	29.42	38.77
LoRA (Rank=1 024)	3.09	18.31	2.58	13.48

来充分训练。(3)优化困难。随着模型参数数量的增加,优化问题变得更加复杂,可能更难找到全局最优解,这可能导致训练过程中的局部最小值,从而影响模型的最终性能。而在相同Rank值下,LoRA的表现远不如GaLore策略,这表明LoRA微调在当前设置下不够有效,需要进一步的优化和调整。

4 案例实证

训练伤防治场景一般具有极高的实时性、精确性、专业性和可操作性,按照武信大模型的信息流程,通过问题输入,利用微调训练的大模型进行推理能力预测输出,并以文本形式反馈给使用者,完成训练伤智能问答。

实证案例1为角色确认“你是谁?”,对比模型为基于传统全参微调方式、GaLore全参微调、LoRA微调得到的3个Qwen2-0.5B基座微调模型,如表7所示。实证结果中,传统全参微调模型的回答中出现了与“你是谁?”不相关的回答。由于传统全参微调涉及对模型所有参数的调整,这可能导致模型过于复杂,从而增加了过拟合的风险,同时训练轮数过多、训练数据量不足等原因也可能导致模型没有足够的信息来学习泛化的规律,而是过度适应了有限数据中的特定样本。相比之下,GaLore全参微调模型与LoRA微调模型的回答均客观正常,效果良好,可能是在训练时通过引入低秩结构减少了模型的复杂度,从而降低了过拟合的风险,提高了泛化能力。

表7 实证案例1(角色确认)

Table 7 Empirical case 1 (role confirmation)

模型	传统全参微调模型	GaLore全参微调模型	LoRA微调模型
实证案例1:你是谁?	颈椎病可以采用中医理疗和按摩针灸进行治疗,效果比较好,平时要注意休息,不要久坐久站,避免风寒,希望对你有所帮助	我是来自阿里云、用不断的学习和进步给你带来便利的超大规模语言模型“通义千问”。	我是阿里云开发的超大规模语言模型,我叫通义千问。

在实证案例2和实证案例3中,训练伤病的具体问题为:“肌肉拉伤怎么办?”“训练中发生骨折怎么办?”GaLore策略全参微调回答的专业性和可操作性较强,每个问题都结合部队训练实际,总结病因并根据不同情况指导治疗,表现出了良好的训练效果。而LoRA微调的回答还具有基础模型的回答范式,表现出一般普遍性。由此可见基于GaLore策略全参微调后的模型能够满足实验预期的效果,极大增强了在训练伤领域回答的专业性和可操作性。

5 结束语

源于训练伤病防治的智能化应用需求,聚焦垂直领域大语言模型训练、评估和系统开发场景,本文提出了武信大模型系统的整体架构,形成了开源数据集+人工数据增强、全参微调训练、多指标能力测评的全链路技术栈,以及数据-模型-训练-应用的开发范式。通过人在回路的数据增强机制提升领域知识密度,采用GaLore全参微调实现参数梯度优化,在降低显存占用的同时增强模型对军事训练伤防治领域知识的表征能力。实验结果验证了GaLore全参微调方法在训练时间、显存占用等方面的性能优势。相关经验和成果可为降低大模型知识幻觉、垂直领域问答大模型的设计与应用提供参考,后续工作将重点突破武信大模型结合知识图谱的检索增强生成能力。

参考文献:

[1] 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. Development and application of ChatGPT large model technology[J]. Journal of Data Acquisition and Processing, 2023,38 (5): 1017-1034.

[2] 崔翛龙,高志强,姬纬通,等.“艾武大模型+”: 一种军事大模型系统的开发与实证[J]. 数据采集与处理,2024,39(3): 588-597.

CUI Xiaolong, GAO Zhiqiang, JI Weitong, et al. "Aiwu Big Model+": Development and demonstration of a military big model system[J]. Journal of Data Acquisition and Processing, 2024, 39 (3): 588-597.

[3] 卡祖铭,赵鹏,张波,等. 面向大语言模型的推荐系统综述[J]. 计算机科学, 2024, 51(S2): 11-21.
KA ZUMING, ZHAO Peng, ZHANG Bo, et al. Overview of recommendation systems for large language models[J]. Computer Science, 2024, 51 (S2): 11-21.

[4] 高志强, 沈佳楠, 姬纬通, 等.大模型技术的军事应用综述[J]. 南京航空航天大学学报, 2024, 56(5): 801-814.
GAO Zhiqiang, SHEN Jianan, JI Weitong, et al. Overview of military application of large model technology[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2024, 56 (5): 801-814.

[5] 任磊,王海腾,董家宝,等. 工业大模型: 体系架构、关键技术与典型应用[J]. 中国科学: 信息科学, 2024, 5 4(11): 2606-2622.
REN Lei, WANG Haiteng, DONG Jiabao, et al. Industrial model: Architecture, key technologies and typical applications[J]. Science in China: Information Science, 2024, 54 (11): 2606-2622.

[6] 张永军,李彦胜,党博,等. 多模态遥感基础大模型: 研究现状与未来展望[J]. 测绘学报, 2024, 53(10): 1942-1954.
ZHANG Yongjun, LI Yansheng, DANG Bo, et al. Multimodal remote sensing basic large model: Research status and future prospects[J]. Journal of Surveying and Mapping, 2024, 53 (10): 1942-1954.

[7] 任芳慧,郭熙铜,彭昕,等. 医疗领域对话系统口语理解综述[J]. 中文信息学报, 2024, 38(1): 24-35.
REN Fanghui, GUO Xitong, PENG Xin, et al. Review of oral comprehension of dialogue system in medical field[J]. Journal of Chinese Information Science, 2024, 38 (1): 24-35.

[8] 汪子健,李传富. 基于BERT的医学智能问答模型研究[J]. 微型电脑应用, 2023, 39(9): 23-25, 29.
WANG Zijian, LI Chuanfu. Research on BERT based medical intelligent question answering model[J]. Microcomputer Application, 2023, 39 (9): 23-25, 29.

[9] AlibabaCloud. QWEN[EB/OL]. (2024-11-30)[2025-05-21]. https://github.com/QwenLM/Qwen.

[10] ZHAO J, ZHANG Z, CHEN B, et al. GaLore: Memory-efficient LLM training by gradient low-rank projection[EB/OL]. (2024-03-06)[2025-03-20]. https: //arxiv.org/abs/2403.03507.

[11] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[EB/OL]. (2021-10-16). https://arxiv. org/abs/2106.09685.

[12] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2021: 4582-4597.

[13] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach, California, USA: PMLR, 2019: 2790-2799.

[14] XU L, LI A, ZHU L, et al. SuperCLUE: A comprehensive Chinese large language model benchmark[EB/OL]. (2025-02-21)[2025-03-20]. https://arxiv.org/pdf/2307.15020.pdf.

[15] 罗文,王厚峰. 大语言模型评测综述[J]. 中文信息学报, 2024, 38(1): 1-23.
LUO Wen, WANG Houfeng. Review of large language model evaluation[J]. Journal of Chinese Information Science, 2024, 38 (1): 1-23.

作者简介:

朱新立(1992-),男,高级工程师,研究方向:大数据与人工智能。

高志强(1989-),**通信作者**,男,副教授,研究方向:军事智能、联邦学习,E-mail: 1090398464@qq.com。

姬纬通(1990-),男,工程师,研究方向:指挥智能化,E-mail:jwt372233354@126.com。

李少华(1992-),男,助教,研究方向:大模型。

李松杰(1998-),男,硕士研究生,研究方向:军事智能。

(编辑:刘彦东)