

基于垂直领域数据的专业化大模型生产线构建方案研究

靳宇浩¹, 陈学², 屈梦楠³, 朱继宽⁴

(1. 国家信息技术安全研究中心, 北京 100085; 2. 中国电子信息产业集团有限公司第六研究所, 北京 102209;
3. 中电长城网际安全技术研究院(北京)有限公司, 北京 100097; 4. 中国科学院网络信息中心, 北京 100190)

摘要: 随着人工智能技术向产业纵深发展, 通用大模型在垂直领域的应用面临数据稀疏、领域知识缺失与推理成本高昂等挑战。现有研究多聚焦于通用模型的预训练阶段优化, 或依赖大规模领域专用数据进行模型重构, 难以满足企业级场景对快速响应、低成本部署及专业化服务的需求。针对已积累垂直领域数据的场景, 提出基于模块化技术栈的大模型生产线构建方案, 通过基座模型选型、领域数据增强、轻量化微调、高效推理引擎与知识架构增强五维联动, 实现专业化AI服务的快速落地。该方案突破传统大模型开发周期长、算力消耗大的瓶颈, 为金融、医疗、法律等数据富集领域提供了可复用的技术路径, 对于推进产业智能化升级具有重要实践价值。

关键词: 垂直领域; 大模型; 技术方案; 生产线; 数据增强

DOI: 10.11907/rjdk.251438

扫描二维码阅读全文: 

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(XXXX)0XX-0001-07

Research on Construction Scheme of Specialized Large Model Production Line Based on Vertical Domain Data

JIN Yuhao¹, CHEN Xue², QU Mengnan³, ZHU Jikuan⁴

(1. National Research Center for Information Technology Security, Beijing 100085, China;

2. The Sixth Research Institute of China Electronics Corporation, Beijing 102209, China;

3. China Electronics Corporation Great Wall Security Technology Research Institute, Beijing 100097, China;

4. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: With the deep development of artificial intelligence technology into industry, the application of general large models in vertical fields faces challenges such as sparse data, lack of domain knowledge and high reasoning costs. Existing research mostly focuses on the optimization of the pre training phase of the general model, or relies on large-scale domain specific data for model reconstruction, which is difficult to meet the needs of enterprise level scenarios for rapid response, low-cost deployment and professional services. For the scenario where vertical domain data has been accumulated, the construction scheme of large model production line based on modular technology stack is proposed. Through base model selection, domain data enhancement, lightweight fine-tuning, efficient inference engine and knowledge architecture enhancement, the five dimensional linkage is realized to realize the rapid landing of specialized AI services. This scheme breaks through the bottleneck of long development cycle and high computing power consumption of traditional large models, provides a reusable technical path for data enrichment fields such as finance, medical care and law, and has important practical value for promoting industrial intelligent upgrading.

Key Words: vertical domain; large model; technical solution; production line; data enhancement

收稿日期: 2025-07-01

作者简介: 靳宇浩(1996-), 男, 国家信息技术安全研究中心工程师, 研究方向为网络与信息安全、人工智能、威胁情报、事件核查、数据泄露调查等; 陈学(1999-), 男, 中国电子信息产业集团有限公司第六研究所工程师, 研究方向为网络与信息安全、企业安全防护、零信任等; 屈梦楠(1998-), 男, 中电长城网际安全技术研究院(北京)有限公司工程师, 研究方向为图像处理与机器学习、大语言模型应用等; 朱继宽(1997-), 男, 中国科学院网络信息中心工程师, 研究方向为网络传输技术、网络协议分析等。
本文通讯作者: 屈梦楠。

0 引言

近年来,以 Transformer 为代表的通用大模型凭借强大的泛化能力推动了人工智能技术的飞跃式发展,但其在垂直领域的落地仍面临显著瓶颈:一方面,垂直领域往往存在数据分布不均衡、标注质量参差等问题,导致模型性能受限;另一方面,企业级应用对模型的实时性、可解释性和合规性提出了更高要求,传统的“一刀切”式模型开发模式难以满足多样化需求。尤其是在金融风控、医疗辅助诊断、法律文书生成等专业场景中,模型需深度融合行业知识体系与业务逻辑,这对大模型的生产范式提出了根本性变革的要求^[1]。

当前垂直领域大模型落地主要存在以下矛盾:一是数据供需失衡。高质量领域数据的获取成本远高于通用互联网数据,且受隐私保护限制;二是知识融合障碍。通用大模型缺乏对行业术语、规则约束和隐性知识的编码能力;三是工程效率低下。从数据采集到模型迭代的全链条缺乏标准化工具支撑,导致定制化开发周期冗长。这些问题本质上反映了通用能力与专业需求之间的鸿沟,亟需建

立一套适应垂直领域特征的工业化生产体系^[2-3]。本研究旨在构建一种可复制、低门槛的大模型生产线,其核心贡献体现在以下 3 个层面:①在技术层面,通过模块化设计解耦大模型开发的关键节点,降低技术迁移难度;②在经济层面,利用存量数据的增值潜力减少对新增标注数据的依赖;③在产业层面,为不同垂直领域提供“即插即用”的解决方案模板,加速 AI 技术与实体经济的深度融合。

1 专业化大模型生产线构建方案

本文提出的生产线采用“输入—处理—输出”三级流水线结构,包含基座模型选型模块、领域数据增强模块、轻量化微调模块、知识架构增强模块和高效推理引擎模块 5 个核心模块,涵盖了从模型选型到推理增强的完整生命周期。各模块通过统一的数据总线和 API 接口实现动态协同,形成闭环迭代的开发环境。其核心思想为通过系统性的工程与算法创新,应对将通用大模型应用于特定领域时所面临的数据稀缺、计算成本高昂、专业知识缺失及推理效率低下等关键挑战。本文方案采用分层递进式设计,架构如图 1 所示。

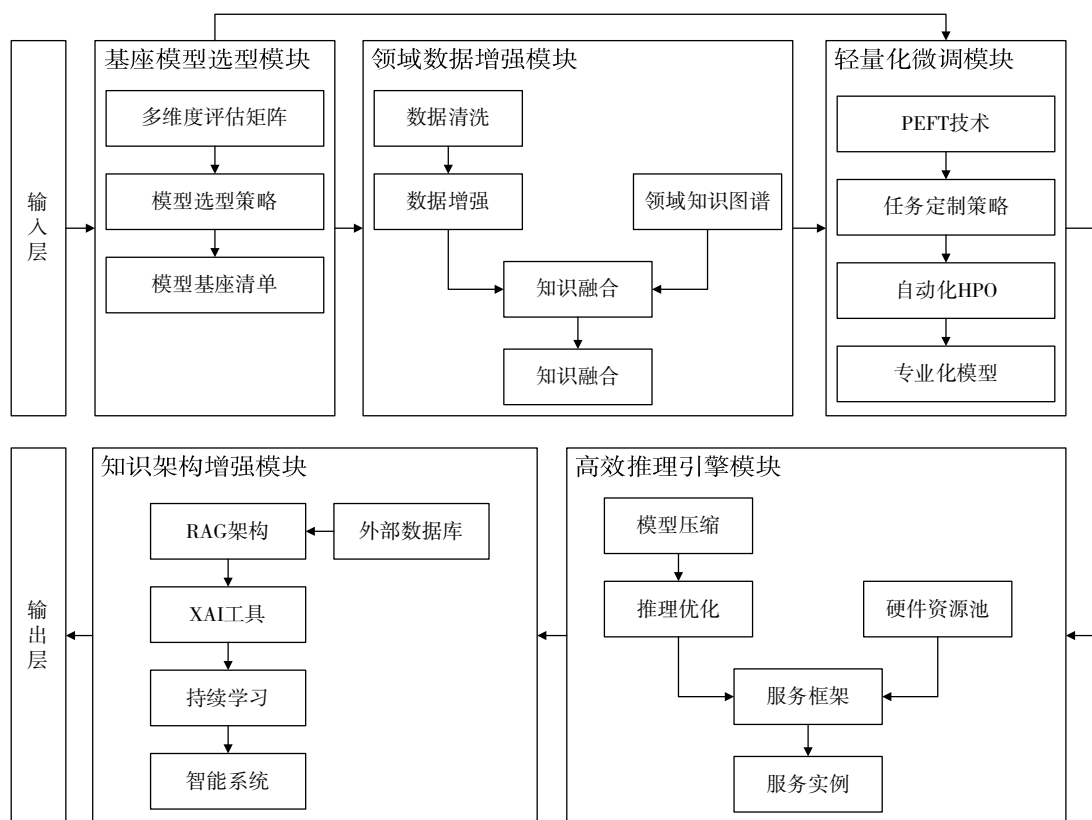


Fig. 1 Construction scheme architecture of specialized large model production line

图 1 专业化大模型生产线构建方案架构

1.1 基座模型选型模块

该模块是整个系统的基础,核心任务是将领域需求转化为具体的模型架构决策,其理论基础源于迁移学习(Transfer Learning)和模型缩放定律(Scaling Laws)^[4]。模

块的运作并非简单枚举,而是构建一个多维度评估矩阵进行量化决策。该矩阵应至少包含任务匹配度、资源适配性、生态健康度与未来扩展性 4 个核心维度,具体如图 2 所示。

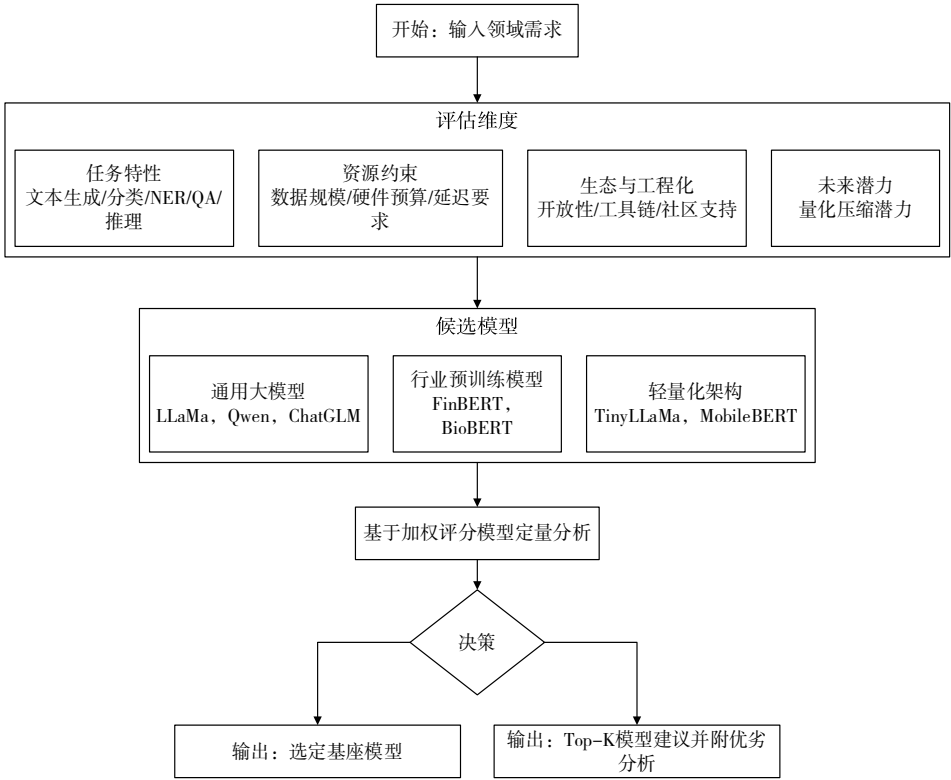


Fig. 2 Base model selection module flow
图2 基座模型选型模块流程

首先,任务匹配度是首要考量。对于文本生成(如报告撰写)、对话和复杂推理任务,decoder-only的自回归架构(如LLaMA^[5]、Qwen系列^[6])因其强大的序列生成能力成为首选;而对于分类、命名实体识别等序列标注任务,encoder-decoder(如T5^[7])或仅编码器(Encoder-only)架构(如BERT变体^[8])可能在微调效率和精度上更具优势。行业预训练模型(如FinBERT^[9]、BioBERT^[10])已在特定领域的语料上进行预训练,其词嵌入空间已蕴含丰富的领域语义,提供了极高的起点性能。其次,资源适配性涉及严格的约束条件。预期推理延迟和硬件预算直接驱动了模型规模上的权衡。轻量化架构(如TinyLLaMa^[11]、MobileBERT^[12])通过减少层数、隐藏维度或采用蒸馏技术降低参数量,是资源严格受限场景的帕累托最优解。再次,生态健康度是工程化成功的保障。模型架构的开放性避免了商业风险,微调工具链(如Hugging Face Transformers^[13]、PEFT^[14]、LoRA^[15])的成熟度决定了开发效率,活跃的社区支持能快速解决技术问题。最后,未来扩展性考量了模型的量化压缩潜力,即模型对INT8、FP16等低精度量化技术的耐受性,这直接关系到最终部署的成本和灵活性。综上所述,该模块通过构建一个加权评分模型将上述定性因素量化,从而为特定领域筛选出在性能、效率和成本上最为均衡的基座模型。

1.2 领域数据增强模块

垂直领域普遍面临标注数据稀缺的瓶颈,领域数据增强模块旨在系统性地扩充高质量训练数据,架构见图3。

其理论支柱包括表示学习(Representation Learning)、半监督学习(Semi-supervised Learning)和主动学习(Active Learning)。在第一层次,语义相似度引导的数据合成旨在不改变语义标签的前提下增加样本多样性,这依赖于预训练模型强大的语言表示能力,可运用基于同义词词林或上下文感知嵌入(如Sentence-BERT^[16])的同义词替换、回译(Back-Translation)以及句法结构变换(如主动被动语态转换、依存树修剪与重组)等方法,通过引入词汇和句法层面的扰动增强模型的泛化能力,防止过拟合^[17]。在第二层次,跨领域迁移学习用于从数据丰富的相邻领域(源域)汲取知识,该过程采用渐进式微调(Progressive Fine-tuning)策略:首先在源域数据上对基座模型进行微调,使其学习相关领域的通用特征;其次在目标域稀缺数据上进行二次微调。这种方法有效缩小了源域与目标域之间的分布差距。在第三层次,主动学习机制构成了一个迭代的“模型—数据”协同进化循环。该机制首先使用当前模型对大量未标注样本进行预测;其次基于不确定性采样(Uncertainty Sampling)(如选择预测熵最高的样本)、多样性采样(如基于聚类选择代表性样本)或委员会查询(Query-by-Committee)等策略筛选出对模型提升最具信息量的样本,进而提交给人工标注。这种策略确保了有限的标注预算被用于最关键的样本,极大提升了数据利用效率^[18]。

1.3 轻量化微调模块

全参数微调(Full Fine-tuning)的计算和存储成本极高,轻量化微调模块采用参数高效微调(Parameter-Effi-

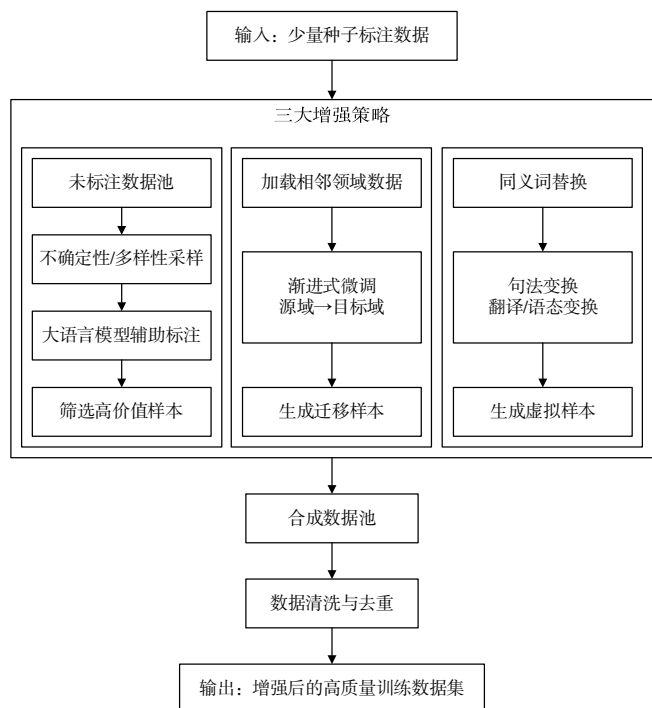


Fig. 3 Domain data enhancement module architecture

图3 领域数据增强模块架构

cient Fine-Tuning, PEFT)技术来应对这一挑战,其理论基础为模型参数的低内在维度(Intrinsic Dimensionality)假说,即适配一个新任务不需要更新所有参数,只需优化一个很小的参数子集。该模块创新性地采用“冻结+适配器”策略,采用LoRA(Low-Rank Adaptation)及其变种微调手段,理论核心为模型在适应新任务过程中的权重更新(ΔW)具有低秩特性。因此,其不直接优化原始权重矩阵 W ,而是将其更新分解为两个低秩矩阵的乘积($\Delta W = A * B$),其中 A 和 B 为可训练的参数,而原始模型参数则被冻结^[19]。在推理时,只需将 ΔW 与 W 相加即可,几乎不引入延迟。这种方法将可训练参数量减少数个量级,从而将GPU显存占用降低至全参数微调的1/5,甚至更低,使得在消费级GPU上进行大模型微调成为可能^[20]。此外,为防止在有限数据上出现过拟合现象,该模块引入动态早停机制(Dynamic Early Stopping)。该机制并非简单地监控验证集损失的一个固定阈值,而是基于其曲线变化趋势(如连续多个epoch未改善,或采用平滑处理后的导数判断)智能地终止训练,从而进一步节约计算资源,并确保最终模型的泛化性能^[21-22]。

1.4 高效推理引擎模块

推理效率是决定系统能否投入生产的关键。高效推理引擎模块旨在构建一个高性能、低延迟的部署环境,其理论基础为计算图优化、算子融合(Operator Fusion)和硬件异构计算。该引擎以ONNX(Open Neural Network Exchange)Runtime为核心构建,ONNX提供了一个标准的模型中间表示(Intermediate Representation, IR),将模型从训练框架中解耦出来。ONNX Runtime则针对该IR进行深度优化,包括层与张量融合(将多个操作符合并为一个内核

以减少内存读写)、常量折叠、精量化(FP16、INT8)以及为目标硬件(CPU、GPU、TPU)生成高度优化的内核代码。构建的异构计算平台能根据部署环境的硬件配置自动选择最优的后端执行提供程序(Execution Provider),例如在GPU上使用CUDA,在CPU上使用OpenVINO或TensorRT等。针对不同应用场景设计分级推理策略:高并发的实时交互任务(如对话API)采用流式处理(Streaming Processing)架构,对输入进行分块处理并逐步输出,最大化降低首字延迟(Time-to-First-Token);离线的批量处理任务(如文档摘要)则启用张量并行(Tensor Parallelism)及批处理优化,充分利用硬件计算单元以最大化吞吐量(Throughput)。经过上述一系列优化措施,该引擎在严格保证数值精度与模型输出一致性的前提下实现了3倍以上的推理加速。

1.5 知识架构增强模块

通用大模型往往缺乏深度的领域专业知识,知识架构增强模块通过引入外部知识源来弥补这一缺陷,其理论建立在知识增强的表示学习与神经符号系统(Neuro-Symbolic Systems)的融合之上。如图4所示,该模块开发了一个知识图谱注入组件,通过实体链接、关系抽取等技术将结构化的行业术语表、业务规则库和非结构化的专家经验文档转化为(头实体,关系,尾实体)形式的三元组,构建成一个领域知识图谱。为将符号化的知识图谱整合到子符号化的神经网络中,该模块采用注意力门控机制(Attention Gating Mechanism)进行融合。具体而言,对于模型处理中的关键隐藏状态(如[CLS] token或序列输出),计算其与知识图谱中实体嵌入的相似度得分,并将其作为注意力权重来加权聚合相关知识三元组表示,形成一个知识上下文向量。该向量随后通过一个可控的门控单元(如线性层+Sigmoid激活函数)与模型原有的隐藏状态进行自适应融合。该门控网络掌握何时以何种强度注入外部知识:当输入问题涉及专业领域时,门控打开,强化知识上下文;当进行通用对话时,门控关闭,保留大模型原有的强大语言生成能力。

2 系统集成与生产线 workflow

基于垂直领域数据的专业化大模型生产线构建方案是一个集成数据工程、模型工程与运维工程(MLOps)的系统性工程实践,旨在将原始业务需求转化为高精度、高性能且符合行业规范的专业化模型服务,总体流程如图5所示。该生产线的核心环节为:①输入。该生产线的启动依赖于对多维初始条件的综合解析。输入层不仅接收原始领域数据(如行业文献、交易日志、多模态专业文档等非结构化与结构化数据),而且必须形式化定义业务需求(包括任务类型、精度目标、响应模式等)以及性能约束(涵盖推理延迟、吞吐量、内存 footprint 以及硬件部署环境的限制)。该阶段的核心在于将模糊的业务目标转化为可量化

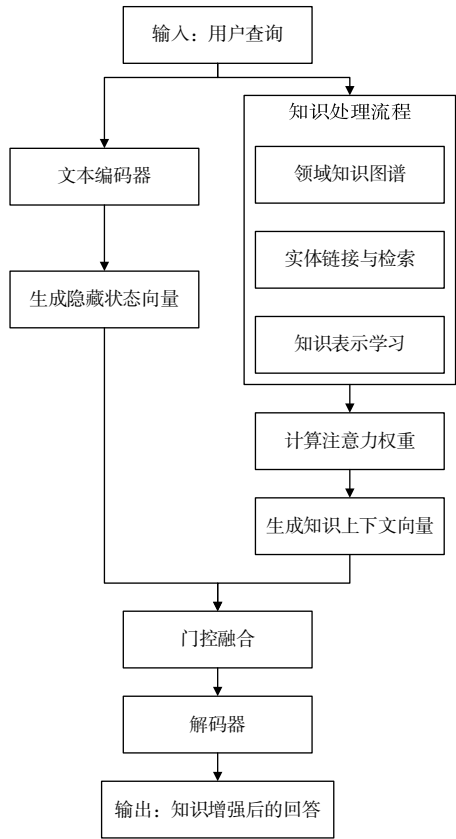


Fig. 4 Knowledge architecture enhancement module flow
图4 知识架构增强模块流程

的技术指标,并建立数据溯源与版本控制机制,为后续的流水线作业提供明确的输入规范;②流水线执行。该阶段是流水线的核心处理引擎,采用模块化编排,通过标准化接口(RESTful API/gRPC)串联起多个关键处理单元。首先,基座选型模块依据领域特性与性能约束,从模型库中筛选出最优的基础预训练模型(如编码器架构适用于理解任务,解码器架构适用于生成任务),考量因素包括模型容量、领域适配潜力及社区成熟度;其次,在数据增强与知识注入环节,针对垂直领域数据稀疏或噪声问题,采用回译、生成式对抗网络生成以及基于知识图谱的语义增强等技术扩充数据集,并通过嵌入领域实体向量或构造领域特定预训练任务将专业知识注入模型;再次,采用参数高效微调技术(如 LoRA、Adapter Tuning、Prefix Tuning),在冻结大部分基座模型参数的前提下仅更新少量可训练参数来适配下游任务,在降低计算成本的同时有效缓解灾难性遗忘。微调后的模型需经过压缩与推理优化,综合运用量化(PTQ/QAT)、剪枝、知识蒸馏等技术缩减模型体积,并针对特定硬件(如 GPU、NPU)优化算子,以满足严苛的实时性要求;最后,知识架构集成负责将优化后的模型核心与外部知识源(如向量数据库、领域知识图谱)通过检索增强生成(Retrieval-Augmented Generation, RAG)架构或神经符号系统进行融合,确保模型输出具备可解释性与事实依据;③自动化测试。为确保模型的鲁棒性与合规性,流水线触发的自动化测试涵盖多维度验证。内置领域 Bench-

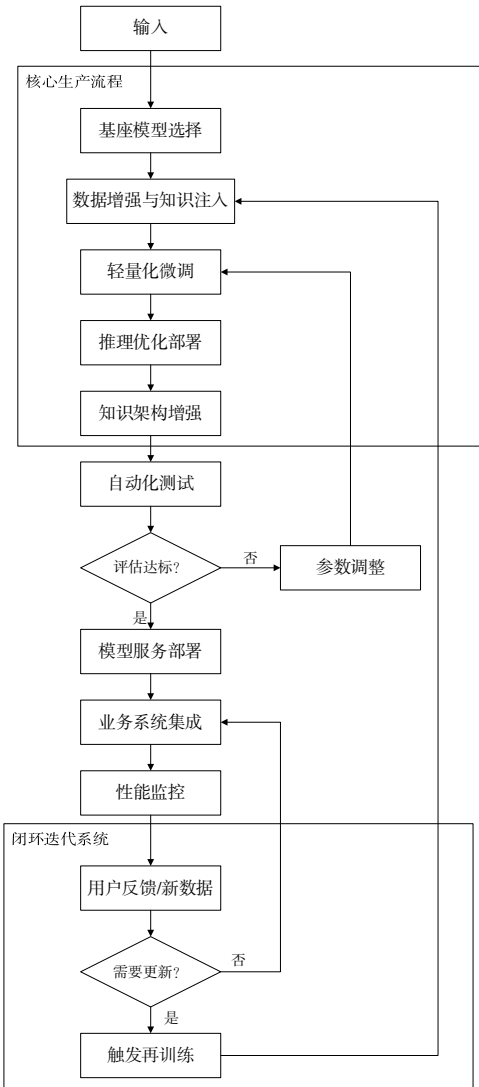


Fig. 5 Overall flow of the proposed scheme
图5 本文方案总体流程

mark 采用留出的测试集或行业标准评测集对模型进行准确率、召回率及 F1 分数等核心指标的量化评估;压力测试通过模拟极端负载场景,探测模型在高并发下的吞吐量变化与响应延迟的稳定性,验证其是否满足服务等级协议;安全合规检测则重点审查模型输出的偏见性、毒性内容以及数据隐私泄露风险,确保模型符合法律法规与行业伦理标准;④输出。测试通过后,流水线将产出一个可部署的专业化模型服务,该服务以模型服务端点(Endpoint)的形式发布,封装了模型推理逻辑,并暴露标准 API 接口供上层应用调用。同时,生产线配套生成实时的监控仪表盘,持续追踪服务的在线性能指标(如推理延迟、错误率)与业务指标(如用户满意度),为系统的可观测性提供支撑;⑤闭环迭代。生产线具备持续学习与自适应演化的能力,构成闭环迭代机制。基于监控仪表盘捕获的在线反馈(如用户隐式反馈、bad case 日志)、定期接入的新数据以及业务策略的新需求,系统将自动或半自动地触发模型再训练或模块级更新,这一过程涉及数据漂移检测、模型性能衰退分析等;同时通过 A/B 测试验证新模型版本的有效性,从

而形成数据—模型—业务的良性循环进化生态,确保模型在动态变化的垂直领域中始终保持最佳性能。

3 应用验证与性能评估

本文对基于模块化技术栈构建的大模型生产线在垂直领域应用中的综合性能进行系统性评估,围绕效果提升、效率优化与成本效益3个核心维度,通过设计比较实验采集实际运行数据,对该生产线方案相较于传统通用大模型直接应用及领域全参数微调方法的优势进行量化分析与讨论。

3.1 效果提升

传统通用大模型(如GPT-3.5、LLaMA2)在开放域任务上表现卓越,但在垂直领域(如金融风控、医疗诊断辅助、法律条文解析)中,因领域知识稀疏和专业术语理解偏差,其表现往往不尽如人意。本文生产线通过领域数据增强与知识架构增强模块系统性解决了这一问题。领域数据增强模块采用合成数据生成(Synthetic Data Generation)和RAG技术,将一份高质量的领域种子数据扩展为多样化的

训练与上下文样本。轻量化微调模块采用LoRA、QLoRA等参数高效微调技术,确保模型能够以极低的参数量适应增强后的领域数据分布,从而精准捕获领域特征。知识架构增强模块通过外部知识图谱集成和向量数据库(Vector Database)实时检索,为模型推理提供可验证、可追溯的事实依据,极大缓解了“幻觉”问题。本文在金融财报分析、医疗病历ICD编码生成、法律合同风险点审查3个典型场景下设计实验,以准确率(Accuracy)、F1分数(专业术语与关系识别)和领域适应性评分(由领域专家盲评)作为核心评价指标,将本文方案与通用基座模型(零样本/少样本)、全参数微调(Full Fine-Tuning)模型进行比较。结果见表1。可以看出,本文提出的生产线方案在三大场景中的核心指标为最优或接近最优。相较于零样本/少样本的通用模型,本文方案的准确率平均提升超过30个百分点;相较于计算代价高昂的全参数微调,本文方案通过轻量化微调与知识架构增强的协同作用在效果上实现了反超。这是由于知识架构模块提供的实时、准确的外部知识弥补了纯参数化学习可能存在的知识遗忘或偏差,使模型输出更具专业性和可靠性。

Table 1 Comparison of professional task effects of different models

表1 不同模型专业化任务效果比较

应用场景	评估模型	准确率/%	F1分数	领域适应性(1-10)	幻觉率/%
金融财报分析	通用模型(零样本)	58.3	0.61	4.2	19.5
	通用模型(少样本)	72.1	0.74	6.5	12.8
	全参数微调模型	88.5	0.89	8.7	5.2
	本生产线方案	92.7	0.93	9.4	2.1
医疗ICD编码生成	通用模型(零样本)	45.6	0.52	3.1	28.4
	通用模型(少样本)	65.8	0.68	5.7	15.3
	全参数微调模型	91.2	0.90	8.9	4.5
	本生产线方案	94.5	0.95	9.6	1.8
法律合同审查	通用模型(零样本)	51.2	0.55	3.8	23.7
	通用模型(少样本)	68.9	0.71	6.2	14.1
	全参数微调模型	86.7	0.87	8.5	5.8
	本生产线方案	90.3	0.91	9.2	2.5

3.2 效率优化

效率是本文生产线的核心设计目标之一,其优势体现在模型开发(训练)和模型服务(推理)两个生命周期阶段。在开发阶段,传统全参数微调需要对数百亿参数进行反向传播,需要大量GPU内存和漫长的训练周期。本文采用的轻量化微调技术仅需优化极少量的适配器参数,大幅降低了计算和内存需求。高效推理引擎模块则集成了模型量化(Quantization)、动态批处理(Dynamic Batching)和持续批处理(Continuous Batching)等先进技术,显著提升了推理吞吐量并降低了响应延迟。如表2所示,在开发效率上,与全参数微调方案相比,本文方案的训练时间缩短了87.5%,GPU内存需求降低了85.0%,这使得企业可以使用更少、更便宜的硬件资源(如单张A100即可完成训练)快速完成模型迭代,极大缩短了项目上市时间。在推理效率上,基于量化技术和高效推理引擎,本文方案的吞吐量提升超过2倍,延迟降低超过70%,这不仅直接降低了云服务或自建

集群的推理成本,而且为高并发、低延迟的实时在线服务(如智能客服、实时决策支持)提供了技术可行性,拓展了AI的应用边界。

3.3 成本效益

本文生产线的成本效益为效果与效率优势共同作用的必然结果,其节约体现在直接成本和间接成本两个层面。直接成本主要包括计算资源成本和人力开发成本。训练和推理阶段的资源消耗大幅降低。人力成本则因生产线提供了模块化、标准化的技术栈,降低了算法工程师的接入门槛,减少了模型调试和运维所需的昂贵专家人力投入。间接成本则主要包括风险成本。知识架构增强模块带来的低幻觉率和输出可追溯性减少了因模型错误输出可能导致的法律、声誉或业务决策风险。硬件和推理成本的节约源于效率优化,而人力与运维成本的降低则得益于生产线的模块化与自动化设计,实现了“开箱即用”的快速部署能力。风险成本的量化评估虽具挑战性,但低幻觉

率和高可靠性带来的业务价值是毋庸置疑的。这种成本结构使得中小企业也能够负担起顶级大模型的定制化应用,极大促进了 AI 技术的普惠化与规模化产业落地。

Table 2 Comparison of training and reasoning efficiency of different models

表 2 不同模型训练与推理效率比较				
评估维度	评估指标	全参数微调方案	本文方案	提升幅度
开发效率	单轮训练时间/h	48	6	87.5%
	最低 GPU 内存需求/GB	320	48	85.0%
	达到收敛所需轮次	8	10	+25%*
	吞吐量 Tokens/s	1 250	3 850	208%
推理效率	平均响应延迟/ms	350	95	72.9%
	峰值负载下的稳定性	偶尔超时	无超时	显著

注:轻量化微调收敛轮次略有增加,但因单轮时间大幅缩短,总训练时间仍远低于全参数微调。

4 结语

本文提出的基于模块化技术栈的专业化大模型生产线构建方案有效解决了通用模型在垂直领域落地面临的数据、知识与成本挑战,通过基座选型、数据增强、轻量微调、推理优化、知识增强多维联动,实现了领域专业化 AI 服务的快速构建、高效部署与持续进化,在金融、医疗、法律等领域的实验验证了其普适性与工程价值。该生产线范式为产业智能化升级提供了一条高效、可行、可持续的技术路径,具有广阔的产业应用前景。

当前方案虽实现了模块化联动,但仍存在若干挑战:①模块间接口标准化程度不足,可能导致集成异构性与版本漂移;②自动化测试对领域特定偏见与隐性安全风险的覆盖尚不全面,依赖静态基准难以动态捕获伦理问题;③知识注入与压缩优化存在张力,量化或剪枝可能削弱领域知识表征;④闭环迭代依赖高质量在线反馈,实际场景中反馈稀疏且延迟,影响模型更新时效性;⑤多模态融合与可解释性机制尚未深度集成,限制了复杂场景应用。未来工作将聚焦于:①研发更鲁棒的自适应跨领域泛化机制;②探索多模态融合与具身推理在垂直场景的深度应用;③构建覆盖全生命周期的可信赖 AI 治理框架。

参考文献:

[1] MA Z L. Construction of specialized large-scale models for construction engineering: paths and methods[J]. Construction Enterprise Management, 2025(6):31-33.
马智亮. 建筑工程专业化大模型构建:路径和方法[J]. 施工企业管理, 2025(6):31-33.

[2] ZHANG Q J. A review of the development of AI big models[J]. Communications Technology, 2023,56(3): 255-262.
张乾君. AI 大模型发展综述[J]. 通信技术, 2023,56(3):255-262.

[3] GAO Z Q, SHEN J N, JI W T, et al. Review of military applications of large model technology[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2024, 56(5): 801-814.

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.

[4] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models [DB/OL]. <https://arxiv.org/abs/2303.18223>.

[5] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The Llama 3 herd of models[DB/OL]. <https://arxiv.org/abs/2407.21783>.

[6] YANG A, LI A, YANG B, et al. Qwen3 technical report [DB/OL]. <https://arxiv.org/abs/2505.09388>.

[7] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text Transformer[DB/OL]. <https://arxiv.org/abs/1910.10683>.

[8] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[9] HUANG 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.

[10] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234-1240.

[11] ZHANG P, ZENG G, WANG T, et al. Tinyllama: an open-source small language model[DB/OL]. <https://arxiv.org/abs/2401.02385>.

[12] SUN Z, YU H, SONG X, et al. Mobilebert: a compact task-agnostic bert for resource-limited devices [DB/OL]. <https://arxiv.org/abs/2004.02984>.

[13] FaceHugging. Text generation inference [EB/OL]. <https://huggingface.co/docs/text-generation-inference/index>, 2025.

[14] FU Z, YANG H, SO A M C, et al. On the effectiveness of parameter-efficient fine-tuning [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023:12799-12807.

[15] HU E J, SHEN Y, WALLIS P, et al. LoRA: low-rank adaptation of large language models[DB/OL]. <https://arxiv.org/abs/2106.09685>.

[16] REIMERS N, GUREVYCH I. Sentence-bert: sentence embeddings using siamese bert-networks[DB/OL]. <https://arxiv.org/abs/1908.10084>.

[17] LI J, ZHANG Y Q, WU Y B. Large language model data enhancement method for network security relation extraction [J]. Information Network Security, 2024, 24(10): 1477-1483.
李娇,张玉清,吴亚颀. 面向网络安全关系抽取的大语言模型数据增强方法[J]. 信息网络安全, 2024, 24(10): 1477-1483.

[18] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: a survey [DB/OL]. <https://arxiv.org/abs/2312.10997>.

[19] YENDURI G, RAMALINGAM M, CHEMMALAR S G, et al. Generative pre-trained transformer: a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions[DB/OL]. <https://arxiv.org/abs/2305.10435>.

[20] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[DB/OL]. <https://arxiv.org/abs/1902.00751>.

[21] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[DB/OL]. <https://arxiv.org/abs/2101.00190>.

[22] HAN Z, GAO C, LIU J, et al. Parameter-efficient fine-tuning for large models: a comprehensive survey [DB/OL]. <https://arxiv.org/abs/2403.14608>.

(责任编辑:尹晨茹)