 DOI: https://doi.org/10.12349/iser.v7i5.10536

Low-Cost Applications of Large Language Models: Mechanisms for Reliable Efficiency Gains and Related Challenges

Hongtao Pan Yexin Gu

28th Research Institute of China Electronics Technology Group Corporation, Nanjing, Jiangsu, 210023, China

Abstract

With the advancement of artificial intelligence technology, the deployment cost of large language models (LLMs) in enterprise software has decreased, providing opportunities for optimizing traditional business processes. This paper focuses on data processing and customer service software, exploring the enabling value of offline LLMs deployment. It is found that offline LLMs can automate data processing, improve customer service efficiency and data security; integrating relevant algorithms and technologies forms a complete data extraction and statistics solution, which can reduce enterprise costs and enhance business accuracy while ensuring data security. The remaining challenges in model accuracy, multi-source data integration and user acceptance, providing a feasible technical approach for low-cost enterprise AI deployment and practical significance for promoting in-depth integration of AI with traditional software.

Keywords

large language models; offline deployment; data processing; customer service; text similarity

AI大模型低成本应用：可靠提效赋能机制与挑战

潘洪涛 顾叶新

中国电子科技集团公司第二十八研究所，中国 · 江苏 南京 210023

摘要

随着人工智能技术发展，AI 大模型在企业软件中部署成本降低，为业务流程优化提供可能。本文针对数据处理与客户服务类软件，探讨 AI 大模型离线部署的赋能价值，可实现数据处理自动化、提升客服效率与数据安全性，结合相关算法及技术形成完整数据提取统计方案,能在保障数据安全的同时节省企业成本,提升业务准确性.鉴于当前模型精度,多源数据融合及用户接受度仍存问题，本文为企业 AI 低成本部署提供一种可行技术路线，对推动 AI 与传统软件深度融合具现实意义。

关键词

AI大模型；离线部署；数据处理；客户服务；文本相似度

1 引言

企业软件应用市场正在经历一场巨大的技术变革.根据IDC发布的《中国AI软件及应用市场研究报告》,2019-2023年中国AI软件市场规模由164.8亿元增长到578.3亿元，年均复合增长率达36.9%。但是，传统的云端AI服务模式在数据安全,成本控制以及响应速度上都存在着较大的问题.麦肯锡2023年全球AI调研报告显示，有67%的企业对云端AI服务的数据泄露表示担忧,有78%的中小企业认为长期使用API进行调用的成本过高[1]。而与此同时，离线AI大模型的发展给解决这些问题带来了希望。将离线AI模型集成进本地软件中,企业可以在保证数据安全的基础上,大大节省开支并且提高工作效率。

目前企业在数据处理以及客户服务两大方面所遇到的问题最为严重,在数据处理方面,传统的人工统计数据的方法效率低下且容易出现人为失误及数据偏差等问题.而在客户服务方面,人工客服的服务速度以及服务的一致性无法满足日益提高的服务水平的要求.因此,本文针对企业软件中AI大模型的离线部署进行研究,在对数据处理类软件以及客户服务类软件中AI大模型的应用进行详细分析的基础上，提出一种可行、有效、低成本的企业级AI应用方案。

2 AI大模型在数据处理类软件中的赋能机制

2.1 离线数据自动统计与分析技术架构

离线数据处理系统利用轻量级的大模型,实现了从数据

【作者简介】潘洪涛（1991–），男，水族，中国贵州福泉人，硕士，工程师，从事软件系统设计研究。

收集到数据分析的全流程自动化.通过模型的轻量化设计,降低部署成本,同时确保性能不受显著影响,使LLMagent能够更广泛地应用于各种实际场景[2]。整个系统是模块化的,主要可由三个部分组成:数据预处理层,智能分析引擎以及结果输出层.数据预处理层用于接收用户上传的Excel,CSV等格式文件，然后根据文件类型进行相应的解析工作,并将这些数据转换成统一的标准格式.智能分析引擎基于已经训练好的统计模型,可以自动地对数据进行清洗、分类统计以及发现其中的问题所在[3]。如果系统发现某些数据存在缺失或者矛盾之处,就会自动标出这些有问题的数据并且给出相应的解决办法,从而大大减少了以往的人工统计过程中出现的各种问题。

表1 离线数据处理系统核心模块功能对比

模块名称	传统处理方式	AI赋能处理方式	效率提升比例	错误率降低比例
数据清洗	人工筛选识别	自动化规则引擎	85%	92%
统计分析	手动计算公式	智能算法处理	78%	89%
异常检测	经验判断	模式识别算法	91%	95%
报表生成	手工制作图表	自动模板匹配	83%	87%

2.2 基于文本相似度的数据提炼算法

文本相似度算法在数据清洗过程中起到重要作用,尤其是对于格式化文档中的结构化数据抽取.该方法利用文本向量之间的余弦相似度来找到需要的数据项。具体来说,先对输入文本做词向量化,再求出目标模式与文本片段之间的相似度。

$$similarity(A,B)=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}}$$

其中A,B是两个文本向量,n是向量维度,这个公式是用两个向量之间的夹角的余弦值来衡量两个文本之间的相似程度，取值范围是[0,1]，越接近1越相似，在实际应用中,系统可设置一个阈值,如果计算出来的结果大于这个阈值就认为是匹配成功并且提取出对应的数据.基于语义理解的匹配方法可以很好地解决不同表达方式带来的问题,保证了数据提取的一致性和准确性。

2.3 OCR驱动的多源数据统一处理方案

多源数据处理方案利用开源OCR技术对扫描图片、PDF文件等非结构化数据进行智能化识别和抽取。系统可以Tesseract OCR为底层识别引擎，同时引入PaddleOCR等深度学习模型提高文字识别准确率.首先对输入图片进行预处理,如去噪,增强对比度,纠正倾斜等操作,然后利用字符分割以及识别方法得到文字内容.对于表格数据识别,系统利用边缘检测方法自动找到表格边框,然后根据行列信息组织数据.该方案对印刷体文档识别准确率,对手写体文档识别准确率均较高，可以满足企业级使用需求。

2.4 数据可视化与趋势预判模型

数据可视化以及趋势预判是数据处理类软件主要功能之一,在以往工作中需要用户自己绘制图表并且根据个人经验对趋势进行预测,这既费时又容易出现由于个人主观因素造成的错误。而AI大模型的离线部署给这个过程带来极大便利,利用自动化的图表生成以及基于机器学习的趋势预判方法大大提高了工作效率以及准确性.对于数据可视化来说,离线AI模型可以根据统计结果的数据特点自动选择最合适的图表形式,比如时序数据会自动生成折线图表示其变化趋势,分类数据会自动生成柱状图或者饼图进行比较,而多维数据则会用散点图或者热力图来显示其相关性,这样一种智能化图表生成方式降低了用户对于图表选择及制作的技术要求。

趋势预判模型是根据历史数据模式识别以及时间序列分析的方法来完成，通过对用户本地的历史数据进行学习,得到数据的变化规律.因为预判是在用户本地的数据上进行计算,不会受到外界数据的影响并且保护用户的隐私,所以更加符合用户的实际需求.这种本地化的趋势预判对于企业的财务预算,库存管理以及业务增长预测等都是非常有帮助的,使企业从被动地查看数据到主动地进行趋势判断并做出相应的决策。

3 AI大模型在客户服务类软件中的应用实践

3.1 离线智能咨询响应系统设计

离线智能咨询系统可基于知识图谱技术,搭建一个完善的问答匹配框架.系统利用预训练语言模型对用户的问题进行语义分析,在本地知识库中找到最相关的内容作为答案.系统具有多轮对话管理功能,可以解决较为复杂的问题.而从响应速度上来看,由于是离线部署,不受网络延迟的影响,平均响应时间远远低于传统人工客服平均响应时间.并且可以支持并发咨询请求,大大缓解了高峰期客服资源紧张的情况。

3.2 知识库构建与匹配优化机制

知识库是分层结构,分为基础问答库,业务规则库以及案例经验库三部分.基础问答库是对于一些常见的问题给出的标准回答,业务规则库是各种业务流程的操作说明,案例经验库是针对一些典型问题给出的解决办法.匹配优化可以使用BM25算法进行初步过滤，然后根据语义相似度进行精确匹配.系统会定期查看用户的咨询记录,发现新的问题及时添加进知识库中。

3.3 咨询记录智能整理与复盘分析

智能记录整理模块利用自然语言处理技术,自动抽取咨

智能记录整理模块利用自然语言处理技术,自动抽取咨询对话中重要信息并形成结构化文档.系统可以判断问题种类,解决情况,客户满意程度等主要因素并根据模板进行记录.复盘分析功能通过对咨询数据进行分析,找出服务提升空间以及知识库完善方向.系统周期性自动生成一份服务质量分析报告，以便更好地提升服务质量。

借助离线AI大模型的深度融合，传统的软件产品在数据处理以及客户服务两个方面都取得了很大的进步,但是模型部署成本,算法准确率以及用户的接受程度等问题还需要进一步解决.而随着硬件的发展以及算法的进步,离线AI的应用将会越来越广泛。

4 低成本部署面临的技术挑战与解决方案

4.1 模型轻量化与性能平衡问题

企业在进行企业软件AI集成时，由于大模型体积大以及有限计算资源之间存在矛盾成为主要问题之一.对于数据处理类软件来说,传统GPT系列模型参数数量达到百亿级别,本地部署需要16GB以上GPU显存，但是企业级应用一般情况下资源有限.利用模型剪枝,知识蒸馏等方法可以将模型参数数量减少,并且不影响主要功能效果.模型压缩率与性能损失之间关系可以用如下公式表示：

$$P_{compressed} = P_{original} \times (1 - \alpha \times \beta)$$

其中 $P_{compressed}$ 是压缩后的模型性能， $P_{original}$ 是原始模型性能，α是压缩率,β是性能下降系数.在压缩率为60%的情况下，性能下降系数小于0.15，保证了数据统计准确率为95%以上[4]。

4.2 离线环境下的模型更新机制

在离线部署环境里,由于不能获取到最新的在线训练数据,所以随着时间推移,模型效果会越来越差.因此,基于增量学习进行本地更新是必要的手段,在本地维护一个用户行为数据缓存池,每隔一段时间进行一次模型参数微调.对于客服类应用,设置问答对质量检测标准,如果用户反馈正确率低于阈值，则自动启动本地模型优化。

模型更新频率以及性能保持之间的平衡是根据用户的使用情况来决定,在频繁使用的情况下,可以每周进行一次参数更新,在较少使用的情况下,可以每月进行一次参数更新.当本地的数据量达到设定数量有效数据时,开始进行模型微调工作,防止小样本导致的过拟合问题.离线更新方式解决了传统软件功能固定的问题，让AI的功能可以随着用户的使用习惯以及业务的发展而变化。

4.3 数据安全与隐私保护策略

离线AI部署的最大好处是数据本地化，但是也必须有相应的安全保障措施.使用端到端加密技术,可以保证用户的数据在模型中被处理的安全性,加密密钥由用户自己生成并保存.对于敏感数据,可以使用差分隐私的方法,在数据统计的结果上加入一定的噪声,从而既保护了个人隐私又不影响统计的效果。

建立数据访问权限分级管理,根据用户角色给予不同功能访问权限,保证核心业务数据只能被授权人员使用.模型推理过程中产生的中间数据用内存加密保存,在计算结束之后马上删除,防止敏感信息泄露.这样可以解决企业用户对于AI应用中数据泄露问题,有利于大模型在金融,医疗等领域应用。

5 结论与展望

本文提出了一种基于AI大模型的企业软件低成本部署的可行技术方案，在保证数据安全性的同时提高应用效率。在数据处理方面，利用文本相似度检测以及OCR技术相结合的方法,可以将非结构化的文档转化为结构化的数据,大大提高了数据统计的准确性和速度.客户服务方面的智能回复,采用本地化部署来保护用户的隐私,同时也为客户提供良好的服务体验。

未来的发展方向是加强多模态融合处理的能力,利用语音,图片等多种形式的数据进行输入,打造更智能的企业应用场景。而模型的小型化以及跨平台的支持将会使AI大模型更容易被中小企业所接受和使用.同时随着边缘计算的发展以及硬件成本的降低，离线AI的应用也会成为企业数字化转型的一个重要推手。

参考文献

[1] 孙力.AI大模型赋能智慧公证视域下法律咨询提效与文书生成适配研究[J].法制博览,2025(29):100-102.
[2] 胡伟，姜晓夏等.大型语言模型智能体：机制和应用的综述[J].指挥信息系统与技术,2024(6):1-11.
[3] 金鑫，谢宇珅等.审计领域大模型应用可信性挑战应对方法[J].指挥信息系统与技术,2025(6):46-53.
[4] 孙常慧.科大讯飞AI转型对创新绩效的影响研究[D].河南财经政法大学,2025.