AI 大语言模型在软件造价领域辅助自动化评估探索

王兵

中国电信股份有限公司重庆分公司，重庆，401147；

摘要：为解决软件造价评估效率低、人力成本高、结果波动大等痛点，本文探索大语言模型（LLM）在该领域的辅助自动化应用。以 NESMA 评估标准为核心，结合企业需求文档质量特征，构建基于 Dify 智能体平台的自动化评估体系，通过 RAG 知识库增强语义识别，设计 “文档预处理 - 功能点分层识别 - 专家校验 - 报告生成” 全流程机制。重庆电信 IT 系统实践表明,该方法使评估效率提升 80%,准确率达人工评估的 80% 以上，年节约人工成本超 100 万元。研究证实 LLM 与专业知识库结合可平衡识别精度与落地可行性，提出多模型融合、动态知识库迭代的未来优化方向，为软件造价智能化评估提供实践参考。

关键词：AI 大语言模型；软件造价；NESMA 标准；自动化评估；RAG 增强；智能体

DOI：10.64216/3080-1508.26.01.042

1 绪论

1.1 研究背景与行业痛点

数字经济深度发展推动软件产业爆发式增长，2023 年国内软件业务收入达 12.32 万亿元，10 年年均增长 16%。软件造价评估作为影响企业数字化响应速度及研发资源配置精度的核心环节，当前主流的 NESMA 功能点分析法存在显著瓶颈：其一，人力依赖度高效率低下，单个中等规模项目评估平均耗时 3.2 人天；其二，评估结果波动性大，专家经验差异导致评估结果离散度达 25%-40%；其三，现有自动化工具适用性不足，现有工具或依赖高度规则化文档适配性不足，或缺乏行业实践积累，识别准确率低于 60%，无法满足实际业务需求。

1.2 研究意义与技术契机

AI 大语言模型在自然语言理解、推理等领域为解决软件造价评估痛点的突破提供新路径，DeepSeek V3、Qwen3 等模型中文语义理解准确率超 91%，专业文档处理场景效率较人工提升 5-10 倍。本研究立足企业需求文档质量现状，探索 LLM 与 NESMA 标准、专家经验的融合方法，构建可落地工具体系，对降低成本、提升决策科学性具有重要意义，其技术框架可为 AI 在其他工程评估领域应用提供参考。

1.3 研究内容与结构

本文先梳理研究现状，再分析 NESMA 评估痛点及 AI 适配点，提出 RAG 增强智能评估方案；阐述系统架构、流程编排及功能实现；通过重庆电信实证案例验证有效性；最后总结成果并提出应用建议与未来方向。

2 相关研究综述

2.1 软件造价评估研究现状

软件造价评估方法分为经验估算法、参数模型法、功能点分析法三类。NESMA 功能点分析法作为国际标准，准确率较前两者提升 20%-30%，成为行业主流。现有研究多聚焦标准优化，如 FV Souto 等结合模糊数学改进模型，将误差降至 15%-20%，但未解决人工效率瓶颈。

2.2 LLM 在专业评估领域的应用进展

AI 在专业文档处理领域已有应用，如李喜梅将 CNN 和知识图谱引入建筑工程量识别，Rahman S. M. Wahidur 通过 RAG 降低合同审查幻觉问题。但软件造价领域研究多处于理论阶段，少数实验未适配实际文档质量，未构建完整流程化工具，本研究填补此空白。

3 问题分析与方案设计

3.1 核心问题实质解析

自动化评估核心目标是实现 “需求文档输入 - 功能点识别 - 工作量计算 - 报告输出” 全流程自动化。关键问题集中于：准确提取 EI/EQ/EO/ILF/EIF 五类功能点元素的语义理解精度；应对文档表述模糊、结构混乱的低质量文档适配能力；整合多环节的流程闭环管控需求。

3.2 技术路径选择

对比两种技术路径：标注大量"需求文档-功能点清单"数据的模型微调模式需 5000 + 高质量标注样本，企业现存样本合格率不足 10%,标注成本超 50 万元；构建 NESMA 规则库、历史案例库的 RAG 增强模式仅需 100

+ 典型案例即可达到基础精度，更符合企业实际。本研究选择路径二，兼顾可行性与经济性。

3.3 整体解决方案框架

以 “AI 能力解耦、流程闭环管控、知识动态更新”为原则，构建 “数据层 - 能力层 - 应用层” 三级架构，实现"文档预处理-智能识别-专家校验-结果输出"全流程自动化。

4 系统设计与实现

4.1 系统架构设计

系统采用分层架构设计，各层松耦合，支持模型快速切换与功能迭代：

1.数据层：包含需求文档库、100 + 已评估项目的历史需求知识库、整合 120 + 条结构化规则的识别规则库。

2.能力层：以 Dify 智能体平台为基座，接入多型号大模型，集成 RAG 增强检索模块，提供可视化编排工具及嵌入模型服务。

3.应用层：涵盖评估任务管理、评价参数管理、功能点识别、专家审核、报告生成五大核心模块。

4.2 核心流程编排设计

遵循"模拟专家思维、分层识别、迭代修正"原则，设计功能点识别流程，具体步骤如下：

1.文档预处理：制定编写规范，按主题拆分章节，单章不超 5 页，通过 Python 脚本自动拆分目录结构，控制模型输入规模。

2.分层识别：按 “系统边界识别→功能类型分类→复杂度判定” 逐层推进，结合 RAG 知识库与 NESMA 规则完成识别。

3.整合修正：通过规则引擎检测分类冲突，自动比对知识库修正。

4.流程闭环：将专家修正结果更新至知识库，实现迭代优化，识别准确率每月提升 4%-5%。

4.3 核心功能实现

4.3.1 评估任务管理与参数配置

实现评估全生命周期管控，支持评估任务创建、分配与归档。参数配置允许预设模板，自定义复杂度权重等参数，适配不同场景，配置效率提升 90%。

4.3.2 功能点识别与校验

实现正向识别与反向校验双功能：正向识别模块依据需求文档 5-10 分钟生成功能点清单及工作量结果；反向校验模块比对第三方清单与需求文档符合性，识别漏报误报，生成差异报告，校验效率提升 8 倍。

4.3.3 专家审核与报告生成

设计"智能预筛+重点审核"机制，系统自动标记低置信度功能点，专家仅需聚焦高风险项审核，审核时间缩短 60%。支持导出含项目信息、功能点明细、专家意见、工作量汇总等内容的 Excel 报告，满足立项审批、成本核算等多场景需求。

4.3.4 历史知识库管理

支持历史评估案例检索与更新，通过相似性算法推荐同类案例，采用增量更新与定期清洗机制，确保知识有效性。

4.4 关键技术难点与解决策略

技术难点	解决策略	实施效果
大文档语义识别精度下降	文档预拆分+章节关联性分析，单批次输入≤5 页，通过上下文向量关联章节信息	识别精度从 65%提升至 80%+
提示词逻辑混乱导致识别误差	按"范围界定→类型判定→复杂度计算"分层设计提示词，嵌入规则校验节点	功能点分类准确率提升 20%
多轮对话级联误差累积	保留关键原始信息（如文档原文片段、规则条款），每轮对话附加历史校验日志	多轮识别误差率控制在 10%以内
精准格式化输出能力不足	智能体编排嵌入 Python 代码节点，对模型输出进行结构化解析与格式修正	输出数据格式化合格率达 100%

5 实证案例分析

5.1 案例背景

选取年均处理 1000+软件开发需求的重庆电信 IT 系统为对象，前期企业采用传统 NESMA 人工评估方式，面临专家资源紧张、评估周期长、结果波动大等问题。2025 年 8-10 月该单位选取 100+典型需求项目采用本研究构建的工具试点，对比工具应用前后的效率、准确率及成本数据。

5.2 应用效果分析

5.2.1 评估效率显著提升

工具应用后单个需求评估耗时从 3.2 人天降至 0.6 人天，效率提升 81.25%，立项高峰期周期从 10 天缩短至 3 天。

5.2.2 评估精度稳步提升

以专家最终审定结果为基准，工具初次识别准确率从试点初期的 50%提升至 83%，专家修正后达 88%以上，结果离散度从 28%降至 12%,波动幅度降低 57.1%。其中，ILF/EIF 类功能点识别准确率最高（90%），EO 类因表述多样性准确率相对较低（82%），后续可通过扩充案例库进一步优化。

5.2.3 成本效益大幅改善

按专家日均人工成本 600 元计算，100 个项目评估直接成本从 19.2 万元降至 3.6 万元，按年均 1000 个项目测算，年节约人工成本超 150 万元。

5.2.4 需求文档质量协同提升

工具对需求文档的结构化要求倒逼业务部门规范文档编写，试点期间需求文档合格率从 10%提升至 65%，模糊表述、信息缺失等问题减少 52%，降低了后续开发需求变更风险。

6 结论与展望

6.1 研究结论

本研究针对软件造价评估的行业痛点，构建了基于 LLM 和 RAG 增强的自动化评估体系，主要结论如下：

1.提出的"RAG 增强+流程编排+专家校验"方案可有效解决低质量文档场景下的自动化评估难题，较传统人工评估效率提升 80%以上，准确率达 88%，实现了效率与精度的平衡。

2.基于 Dify 智能体平台的架构设计实现了 AI 能力与业务流程的解耦，支持多模型切换及知识库增量更新，适配不同企业的个性化需求，落地性强。

3.文档预处理、分层提示词设计、多轮对话误差控制等关键技术策略，有效提升了 LLM 在专业场景的应用精度，为同类研究提供了技术参考。

4.重庆电信的实证案例表明，该方法可显著降低评估成本、提升决策质量，同时带动需求文档质量协同改善，具备良好的经济效益。

6.2 实际应用建议

基于研究成果，对企业应用 LLM 辅助软件造价评估提出以下建议：

1.分阶段推广落地：初期选取标准化程度高的项目类型（如简单二次开发）试点，积累案例数据优化知识库后，再推广至复杂定制开发项目；

2.构建系统知识库：根据软件系统大类特点建立系统分类知识库，解决系统特征沉淀问题，加速模型适配；

3.建立专家反馈机制：将专家审核过程转化为知识标注流程，通过"识别-审核-标注-更新"闭环持续优化知识库，提升模型自主识别能力；

4.规范需求文档管理：配套制定需求文档编写规范及审核标准，从源头提升文档质量，降低模型识别难度。

6.3 未来研究方向

本研究仍存在进一步优化的空间，未来可从以下方向深化：

（1）多模型融合优化：引入计算机视觉技术处理文档中的表格、流程图等非文本信息，结合 LLM 实现多模态信息融合识别，提升复杂文档处理能力；

（2）动态知识库构建：基于知识图谱技术构建动态更新的 NESMA 规则库，结合实时行业案例数据，实现规则的自动演进与适配；

（3）端到端自动化升级：通过强化学习训练评估模型，减少对专家审核的依赖，实现"需求输入-工作量输出"的端到端自动化。

参考文献

[1]中国软件行业协会.(2024).中国软件产业高质量发展报告.中国软件行业协会.2024

[2]蚂蚁开源技术增长团队 & Inclusion AI. (2025).中文开源大模型全景报告. 蚂蚁集团.2025

[3]FV Souto. (2014). COSMIC Approximate Sizing Using a Fuzzy Logic Approach. IEEE IWSM.Mensura. 10.1109/IWSM.Mensura.2014.44

[4]李喜梅.(2021).基于人工智能技术的建筑工程造价估算研究.城市建筑.2021.02:146-148

[5]Rahman, S. M. W. (2025). Legal Query RAG (LQ-RAG). IEEE Access, 13, 10887211

作者简介：王兵（1976.12—），男，汉族，四川广安，本科，高级工程师，大模型 IT 领域应用，中国电信股份有限公司重庆分公司。