Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(6), 1361-1371
Published Online November 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.146127

Hans汉斯

基于大语言模型的财务报告指标抽取智能体方法

张亚豪[1]，施水才[1,2]，王洪俊[1,2]，秦　疆[1]

[1]北京信息科技大学计算机学院，北京
[2]拓尔思信息技术股份有限公司，北京

收稿日期：2025年9月19日；录用日期：2025年10月29日；发布日期：2025年11月7日

摘　要

金融年报中的信息抽取因其复杂的PDF格式和超长上下文而极具挑战。传统的检索增强生成(RAG)方法受限于单步、静态的检索范式，一旦初始查询与文档表述不匹配，便容易失败。为解决这一检索脆弱性问题，本文提出了一个名为LedgerLens的多智能体协作框架。该框架借鉴人类分析师的认知模式，其核心研究智能体(Researcher Agent)通过“检索－分析－精炼”的迭代循环，在初步检索结果不佳时能够自主重构查询并进行多轮尝试，直至精准定位目标信息。在自建的银行年报问答数据集(BAR-QA)上的实验结果表明，LedgerLens在指标抽取任务中取得了94.1%的F1分数，并在大多数任务上实现了领先表现。研究结果证明，引入基于智能体的迭代查询优化机制，是突破传统RAG在复杂真实场景中检索瓶颈的有效途径。

关键词

大语言模型，智能体框架，检索增强生成，财务报告分析

Agent-Based Financial Report Indicator Extraction with Large Language Models

Yahao Zhang[1], Shuicai Shi[1,2], Hongjun Wang[1,2], Jiang Qin[1]

[1]College of Computer Science, Beijing Information Science and Technology University, Beijing
[2]TRS, Beijing

Received: September 19, 2025; accepted: October 29, 2025; published: November 7, 2025

Abstract

Information extraction from financial annual reports is highly challenging due to their complex PDF

文章引用: 张亚豪, 施水才, 王洪俊, 秦疆. 基于大语言模型的财务报告指标抽取智能体方法[J]. 人工智能与机器人研究, 2025, 14(6): 1361-1371. DOI: 10.12677/airr.2025.146127

formatting and extremely long contexts. Traditional retrieval-augmented generation (RAG) methods rely on a single-step, static retrieval paradigm, which is prone to failure when the initial query does not align with document expressions. To address this retrieval vulnerability, this study proposes LedgerLens, a multi-agent collaborative framework. Inspired by the cognitive process of human analysts, the core Researcher Agent operates through an iterative "retrieve-nalyze-refine" loop, enabling it to autonomously reformulate queries and conduct multiple retrieval attempts until the target information is accurately located. Experiments on a self-constructed Bank Annual Report Question Answering dataset (BAR-QA) demonstrate that LedgerLens achieves an F1 score of 94.1% in the indicator extraction task and outperforms baselines on most tasks. These results indicate that introducing an agent-based iterative query optimization mechanism provides an effective solution to overcoming the retrieval bottlenecks of traditional RAG in complex real-world scenarios.

Keywords

Large Language Models (LLMs), Agent Framework, Retrieval-Augmented Generation (RAG), Financial Report Analysis

1. 引言

银行年报是投资分析与监管审计的核心文档，包含大量结构化指标(如净利润、资产总额等)。这些信息通常以 PDF 表格的形式发布，排版复杂、格式多样，给自动化抽取任务带来了巨大挑战。传统方法(如基于规则或模板的方法)通常缺乏鲁棒性，而单轮 RAG 搜索虽能缓解幻觉问题，但仍受限于上下文覆盖与检索准确性。

为提升复杂文档中的信息匹配能力，深度研究智能体(Deep Research Agents, DRAs)提供了一个强大的新范式。与单轮问答不同，DRAs 通过模仿人类研究员的“检索–分析–精炼”(Plan-Act-Reflect)循环，支持通过多轮迭代的流程来处理复杂任务[1]。

Huang 等人在其系统性综述中，明确指出了智能体在任务分解、多轮交互和工具调用方面的巨大潜力[1]。在实践中，Zheng 等人提出的 DeepResearcher 框架，通过强化学习来训练智能体自适应地优化查询策略，相关实验证据表明，该方法可提高决策链的稳定性，在检索覆盖率和答案生成质量上分别提升了约 28.9%和 7.2% [2]。

在银行年报的指标抽取这一具体任务中，观察到，多数指标的值并不需要复杂的跨片段逻辑推理，而是能够明确地定位到 PDF 中特定表格的某一单一来源。因此，本文认为，直接引入为开放域问答设计的多跳(multi-hop)或图结构检索机制(如 HopRAG)可能并非最优解，可能造成系统复杂度与工程维护成本的增加。

基于上述观察，本文提出了一种更加精简、高效且任务导向的智能体架构。本文的流程由三个协同工作的 Agent 角色构成：

规划智能体(Planner Agent)：作为任务的起点，负责将用户模糊的自然语言查询解析并规范化为一个精确的、结构化的任务指令，为后续的精准检索提供依据。

研究智能体(Researcher Agent)：作为框架的核心，负责执行一个“检索–分析–精炼”的迭代循环。当初步检索结果不满足要求时，它能自主反思并优化查询，发起新一轮的检索，直至定位到最高质量的

信息源。

回答智能体(Answerer Agent)：作为任务的终点，负责从研究智能体提供的最佳信息块中，精准地抽取出指标的“数值”和“单位”，并附上来源信息，确保结果的完全可追溯性。

该整体流程既保留了深度研究智能体(DRA)多轮反思与迭代的核心优势，又充分贴合了金融指标抽取任务的结构化特性，避免了因设计冗余导致的额外复杂性。本文的主要贡献在于，本文展示了如何将一个通用的、强大的 Agent 范式，巧妙地适配并优化于一个具体的、有重要价值的金融场景中，从而在保证高精度的同时，也兼顾了系统的可控性与效率。

2. 相关工作

本文的工作 LedgerLens 旨在通过智能体(Agent)框架解决长篇、复杂金融文档中的信息抽取问题。这项研究与三个紧密相关的领域交叉：1) 金融领域的检索增强生成(RAG)技术；2) 长文档与多模态金融问答的特定挑战；3) 作为新兴解决方案的深度研究智能体(DRA)范式。

2.1. 金融领域的检索增强生成(RAG)

检索增强生成(RAG)已成为将大型语言模型应用于知识密集型领域的事实标准[3]。在金融领域，构建一个高效的 RAG 系统是众多研究的核心。Wang 等人基于 LLM-RAG 构建了一个智能金融数据分析系统，展示了其应用潜力[4]。

研究者们致力于优化 RAG 流程的各个环节，尤其是在检索和重排序阶段。

在检索端，为了提升召回率和精准度，Lee 等人专注于改进金融文档问答模型的检索策略，而 Kim 等人则系统性地优化了不同检索策略的组合[5] [6]。更有研究者探索了将知识图谱(Knowledge Graphs)与传统向量检索相结合的混合方法，以期利用结构化知识增强检索效果[3]。

在排序端，由于初步检索召回的文档可能存在噪声，精细化的重排序(Re-ranking)至关重要。Lee 等人在 FinanceRAG 挑战中，通过一个 Multi-Reranker 模块最大化了 RAG 系统的性能，证明了先进重排序策略的有效性[7]。

这些工作极大地提升了 RAG 流程在金融任务上的性能，但它们大多仍在一个相对固定的“检索–排序–生成”流程内进行优化。

2.2. 长文档与多模态金融问答的挑战

将 RAG 应用于真实的金融场景，必须解决金融文档本身带来的两大固有挑战：超长的上下文和多模态的内容。

长文档处理与分块(Long Document Processing & Chunking)：金融年报等文档通常长达数百页，远超多数模型的上下文窗口[8]。Reddy 等人推出的 DocFinQA 基准，首次将完整的长文档上下文引入金融问答任务，极大地推动了该领域的发展[8]。如何有效地将这些长文档切分成有意义的、信息完整的文本块(Chunks)成为一个关键的前置任务[9]。Wang 等人和 Jimeno-Yepes 等人分别对此进行了深入研究，探索了不同的分块策略对下游 RAG 性能的影响[10]。

多模态信息处理(Multi-modal Information Processing)：财务报告不仅包含文本，还有大量的表格和图表，这些是关键信息的富集区[11] [12]。传统的纯文本 RAG 无法处理这些视觉信息。为此，Jiang 等人探索了以图像为中心、利用 LLM 分析图表的多模态 RAG 方法[11]。Gondhalekar 等人则提出了一个优化的多模态 RAG 框架 MultiFinRAG，旨在更全面地整合金融文档中的文本与视觉元素[12]。

此外，高质量的基准数据集对于推动领域发展至关重要。除了 DocFinQA，Islam 等人早前发布的 FinanceBench 为金融文档理解提供了一个开放的基准[9]，而 Nguyen 等人近期构建的 SEC-QA 则为金融

问答提供了一个新的系统性评测语料库[13]。这些工作共同描绘了金融文档分析领域的复杂性和挑战性。

2.3. 深度研究智能体 (DRA) 作为新范式

为了从根本上解决传统 RAG 流程固定、缺乏灵活性的问题，深度研究智能体(DRA)这一新范式应运而生[1]。与按部就班执行任务的 RAG 系统不同，DRA 旨在模仿人类研究员的认知过程，通过自主的检索、分析和精炼来解决复杂问题。

Huang 等人在其综述性工作中，将 DRA 的核心架构概括为一个“检索 - 分析 - 精炼”的循环，并系统地分析了其在任务分解、工具调用等方面的能力[1]。这种 Agentic 方法不再将检索和生成视为孤立的步骤，而是将其作为 Agent 在解决问题过程中可以调用的“工具”。

本文的工作 LedgerLens 正是建立在这一前沿思想之上。本文认为，面对金融年报这种半结构化的复杂文档，与其不断为固定的 RAG 流程增加更多的组件(如多重检索器、重排序器)，不如转向一个更灵活的智能体框架。LedgerLens 通过将任务分解给不同角色的 Agent (Planner, Researcher, Answerer)，并赋予 Researcher Agent 进行多轮迭代、动态调整查询策略的能力，这是首次将 DRA 的通用思想有针对性地应用于金融指标抽取任务。

本文研究方法主要分为五个阶段，涵盖从系统构建到数据处理与异常识别的全过程，具体包括设备研究与软件开发、数据采集与预处理、异常数据检测、图像模型辅助判定、结果分析与方法评估。设备研究与软件开发部分，自主研发并设计了非接触式磁记忆检测设备，搭配开发具有采集磁传感器数据、实时定位与同步图像记录功能的采集软件。数据采集与预处理部分，在模拟工况环境中布置多组外部干扰源，利用所开发系统同步采集磁信号、图像信息及位置信息。采集完成后，基于经纬度信息计算对应采集距离，并将磁信号按距离窗口进行均值处理，以降低数据波动对后续分析的影响。在异常数据检测阶段，利用孤立森林与局部分析算法对处理后的磁信号进行无监督建模，识别序列中显著偏离整体分布的异常点与局部突变的异常点，作为潜在缺陷或外部干扰的候选位置。模型验证与数据分析部分，引入基于 YOLO 的图像分类模型对采集图像进行辅助识别。结合图像结果分析磁信号异常点对应位置是否存在外部干扰物，从而排除由外部干扰引发的误判，提升缺陷识别的可靠性。结果分析部分，综合磁信号异常点与图像分类结果，对模型辅助筛查结果进行统计分析与人工验证，评估所提出方法在实验环境下的识别准确率与应用可行性。

3. 方法

针对从超长、半结构化的银行年报中精准抽取财务指标的挑战，本文设计并实现了一个名为 LedgerLens 的多智能体(Multi-Agent)协作框架。本章将详细介绍 LedgerLens 的总体架构、文档预处理流程，以及其核心的 Agentic 工作流。

3.1. 总体架构

LedgerLens 的设计思想源于深度研究智能体(Deep Research Agents, DRAs)的前沿范式，即通过模仿人类研究员的“检索 - 分析 - 精炼”循环来解决复杂问题[1]。与通用的 DRA 框架不同，LedgerLens 针对金融指标抽取的“精定位”特性进行了深度优化，形成了一个由规划、研究和回答三个核心智能体高效协作的精简流程。

LedgerLens 的整体工作流程如图 1 所示。首先，用户的自然语言查询被规划智能体解析为结构化的任务指令。随后，研究智能体根据该指令，通过一个“检索 - 分析 - 精炼”的迭代循环，在知识库中进行信息的精准定位与筛选。最后，回答智能体从研究智能体提供的最优信息中提取并输出结构化的答案。

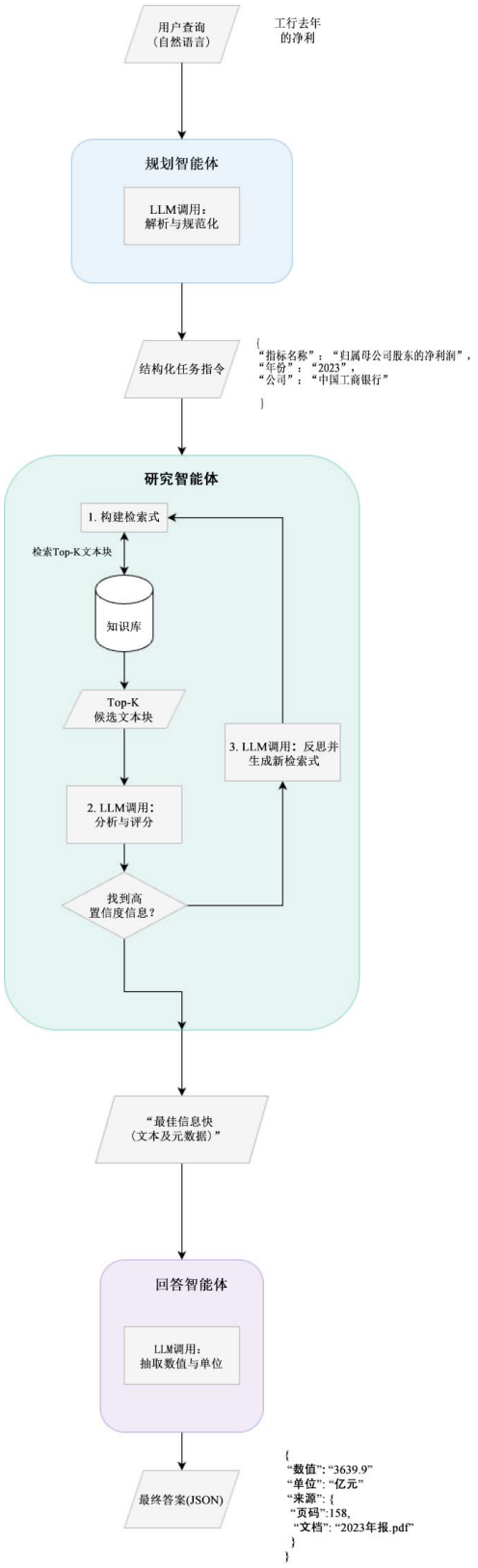

Figure 1. Data architecture diagram
图 1. 数据架构图

3.2. 文档预处理与知识库构建

高质量的知识库是 LedgerLens 高效工作的基础。本文的预处理流程包含以下三个关键步骤：

PDF 的结构化解析：本文采用 monkeyCOR 解析工具，将原始的 PDF 年报直接转换为 Markdown 格式的文本。相比于纯文本，Markdown 格式能够更好地保留文档的层级结构(如各级标题)和表格的行列信息，这为后续的结构化分块奠定了基础。

结构化分块(Structural Chunking)：为了在保持信息完整性和控制分块粒度之间取得平衡，本文摒弃了单一的分块方法，转而采用一种以语义为导向的混合分块策略：

以标题为边界：首先，本文将整份文档按照其自然的章节、子章节等各级标题进行宏观分割。所有的分块操作都在一个末级标题所限定的语义边界内进行。

以段落为单位：在每个标题边界内，段落是构成文本块(Chunk)的基本单位。本文优先将每一个自然段落作为一个独立的 Chunk。这种方式最大程度地保留了文本的语义完整性。

长段落的滑动窗口切分：对于少数极长的段落(例如，超过 1024 个 token)，为了避免超出模型处理能力，本文会启动一个回退(fallback)机制：使用一个固定长度的滑动窗口(例如，512 个 token 的窗口长度，带有 10%的重叠)对其进行切分。重叠部分确保了切分边界不会丢失关键的上下文信息。

表格独立分块：所有的表格都被识别并提取为独立的、特殊的表格块，与文本块分开处理。

表格的针对性向量化策略：表格是财务指标的核心载体，但其包含的大量数字会对基于语义的向量检索造成干扰。为了解决这个问题，本文在向量化表格块时，采用了一种创新的策略：在生成文本嵌入(Text Embedding)之前，本文从表格的 Markdown 文本中，临时移除了所有的数字字符以及 Markdown 的表格格式符(如，-)。这样做能够让向量模型专注于学习表格的结构和标题等元信息的语义(例如，“合并资产负债表”、“主要会计数据摘要”)，而不是具体的数值。这使得 Researcher Agent 能够更准确地根据语义召回“哪个表格是关于什么的”，显著提升了检索的准确率。

3.3. LedgerLens Agentic 工作流

LedgerLens 的核心是一个由四个 Agent 角色组成的协作链，每个角色各司其职，共同完成精准的指标抽取任务。

Planner Agent 是工作流的起点。它的核心职责是将用户输入的、模糊的自然语言查询，转化为一个精确、无歧义的结构化任务指令。这是规划阶段的核心。例如，当接收到查询“工行去年的净利是多少”时，Planner Agent 会输出一个 JSON 对象。

这种结构化的指令极大地降低了后续 Agent 的工作难度，避免了因自然语言歧义造成的检索错误。

研究智能体是信息检索与验证的核心，它将传统 RAG 中分离的“检索”和“重排”步骤，融合进一个由 Agent 主导的、本文称之为“检索–分析–精炼”(Retrieve-Analyze-Refine)的循环中：

检索(Retrieve)：接收到规划智能体的结构化指令后，研究智能体首先构建初始检索式。它会调用其工具箱(Toolbox)中的工具来执行检索。该工具箱主要包含：

Search_Tool：负责在向量数据库中执行核心的语义相似度检索。

Directory_Tool：负责利用文档的目录层级结构进行宏观导航。对于需要跨章节的宽泛问题，Agent 可优先调用此工具来定位到具体的章节，缩小后续语义检索的范围。

在第一轮检索中，Agent 通常会调用 Search_Tool，召回一个初步的候选文本块列表(如 Top-20)。

分析(Analyze)：这是与传统 RAG 最根本的区别。研究智能体会调用 LLM 对每一个候选块进行相关性评估和打分(例如，从 1 到 10 分)，以判断其与任务指令的契合度。

决策(Decide)：Agent 会根据分析阶段的打分结果执行决策：

如果存在高分(如≥9 分)的候选块，Agent 判定已找到足够好的信息源(“最优信息块”)，便会将这个最高分的文本块传递给回答智能体，并结束研究循环。

如果所有候选块得分均不高，Agent 则进入下一步的“反思”阶段。

精炼(Refine)：在反思阶段，Agent 会调用 LLM 分析当前结果不佳的原因(例如，关键词太宽泛、缺少上下文限定等)，并自主生成一个更精确的新检索式。随后，Agent 将基于新的检索式重新进入检索循环，开始新一轮的迭代。

这个迭代循环确保了研究智能体不仅执行检索任务，还具备自主反思与调整能力。它通过自我评估和动态调整，层层深入，直至锁定最精准的信息源。

Answerer Agent 是工作流的终点，负责将 Researcher Agent 筛选出的“金块”(golden chunk)转化为最终交付物。它接收排序最高的、最相关的单一文本块，并执行两个最终动作：

精确抽取：通过一个带有抽取指令的 Prompt，引导 LLM 从文本块中提取出指标的“数值”和“单位”(例如，“19,187,433”和“百万元”)。

提供溯源：在输出最终答案的同时，附上该信息所在的源文档页码、表格或行号。这在金融审计等严肃场景中至关重要，确保了所有结果都是可验证、可追溯的。

4. 实验

为了系统地评估 LedgerLens 框架，本文设计了一系列实验。本文的评估核心(4.5.1 节)聚焦于其在银行年报财务指标抽取这一主要任务上的表现。随后，通过一系列补充实验，包括对复杂问答任务的泛化能力分析(4.5.2 节)、验证框架各组件有效性的消融实验(4.5.3 节)，以及在公开基准上的对标和效率分析(4.5.4 节)，来对其进行全方位画像。

4.1. 数据集

FinanceBench：为了确保工作的可比性和权威性，选取了 FinanceBench 作为的公开评测基准[9]。该数据集由 Islam 等人创建，已被 FinSage [14]等近期 SOTA 工作用作其核心性能评估基准[9] [10]。在 FinanceBench 上进行实验，能够让 LedgerLens 直接与这些已发表的先进成果在公平的条件下进行性能对标。

银行年报问答数据集(BAR-QA)：考虑到公开数据集无法完全覆盖银行年报中特有的复杂性和指标抽取需求，自行构建并标注了一个高质量的银行年报问答数据集。该数据集包含 50 份国内主要商业银行的年度报告，并针对这些报告设计了 500 个问答对。所有问答对均由两位金融领域的专业人士独立进行“背对背”标注，对于不一致的样本，再由第三位资深专家进行最终仲裁。经过计算，本文的数据集标注者间一致性(Inter-Annotator Agreement)的 Cohen’s Kappa 系数达到了 0.92，表明了其高质量和可靠性。

4.2. 对比模型

为了全面验证 LedgerLens 的性能，本文将其与两类基线模型进行了对比：亲手复现的 SOTA 模型，以及直接引用自相关文献的 SOTA 模型的报告结果。

DocFinQA-style Retriever：该基线模型复现了 DocFinQA (Reddy et al., 2024)论文中被验证为最有效的检索增强问答流程[8]。该流程首先将长篇年报切分为 512 个 token 的有重叠文本块，并使用 BGE-large-en-v1.5 模型进行向量化。在接收到查询后，它使用向量相似度检索 Top-5 最相关的文本块，并将其与原始问题拼接后输入 GPT-4o 生成最终答案。

FinSage-style RAG：该基线模型复现了 FinSage 论文中提出的、更先进的多路径 RAG 框架的核心思想[14]。该模型并行使用 BM25 稀疏检索和 BGE-large-en-v1.5 稠密检索来召回候选文本块。随后，它引

入一个 bge-reranker-large 交叉编码器模型对候选块进行精准打分和重排序，选取最优的 Top-5 片段送入 GPT-4o 进行答案生成。

4.3. 评估指标(Evaluation Metrics)

根据任务类型的不同，采用两套评估指标：

指标抽取任务(主要任务)：对于 BAR-QA 数据集中的指标抽取任务，采用信息抽取领域的标准指标：精确率(Precision)，召回率(Recall)，和 F1 分数(F1-Score)。一个指标被认为抽取正确，当且仅当其“数值”和“单位”均与标准答案完全匹配。

问答任务(泛化能力及公开基准)：对于涉及复杂推理的问答和 FinanceBench 上的任务，采用准确率(Accuracy)，即模型生成的最终答案与标准答案完全一致的比例

4.4. 实验设置

为确保对比的公平性，所有对比模型(LedgerLens, DocFinQA-style, FinSage-style)在实验中遵循了严格的控制变量原则，且防止出现偶然现象，以下实验均是 10 次实验结果的平均值作为最终结果：

(1) 统一的核心引擎：所有模型的“大脑”或“生成器”均统一采用 GPT-4o 模型，通过 API 进行调用(temperature = 0.1)。这确保了性能的差异主要来源于框架设计，而非底层语言模型的不同

(2) 统一的嵌入与检索工具：所有模型的向量化均采用 BGE-large-en-v1.5 嵌入模型。所有向量检索均通过 FAISS 实现。所有涉及重排序的模块，均统一使用 bge-reranker-large 交叉编码器。

(3) 具体的参数设置：检索 Top-K：在不使用重排序器的 DocFinQA-style 模型中，直接检索 Top-5 的文本块。在包含重排序的 FinSage-style 和 LedgerLens 中，首先召回一个包含 Top-20 的候选池，再由重排序器精选出 Top-5。Agent 提示词设计：LedgerLens 中各 Agent 的提示词经过精心设计。例如，PlannerAgent 的提示词会引导模型将用户输入(如“工商银行去年的净利润”)解析为一个包含{indicator：“归属母公司股东的净利润”，year：“2023”，company：“工商银行”}等字段的结构化 JSON。Researcher Agent 的提示词则鼓励其进行多轮迭代搜索，并在后续轮次中主动加入“表格标题”、“货币单位”等关键词来优化查询。相关性阈值：在 LedgerLens 的 Researcher Agent 进行多轮检索时，设定了一个余弦相似度阈值 0.85。若一轮检索返回的最优文本块与上一轮的最优块相似度低于此阈值，Agent 会判断可能已偏离主题，并触发“反思”机制来调整查询策略。

(4) 硬件环境：所有实验均在配置有 4 张 NVIDIA A100 (80GB) GPU 的服务器上进行。

4.5. 实验结果与分析

4.5.1. 主要任务表现：指标抽取

在 BAR-QA 数据集上，针对核心的指标抽取任务，对各模型进行了评测。

Table 1. Performance of various models on the BAR-QA metric extraction task (%)
表 1. 各模型在 BAR-QA 指标抽取任务上的表现(%)

模型(Model)	精确率(P)	召回率(R)	F1 分数
DocFinQA-style Retriever	81.3	76.5	78.8
FinSage-style RAG	89.5	86.2	87.8
LedgerLens	95.2	93.0	94.1

表 1 的结果清晰地展示了 LedgerLens 在核心任务上的绝对优势。其 F1 分数达到了 94.1%，显著高于 FinSage-style 的 87.8%。本文分析这主要归功于 Planner Agent 对查询的规范化，以及 Researcher Agent

的多轮迭代搜索策略，这两者共同确保了能更大概率、更精准地定位到包含正确指标的唯一来源。

4.5.2. 泛化能力分析：处理复杂问答任务

尽管 LedgerLens 主要为指标抽取设计，本文仍希望测试其框架在处理更复杂的、需要推理的问答任务上的泛化能力，因此从 BAR-QA 中选取了需要跨章节和复杂计算的子集进行测试。

Table 2. Performance of various models on the BAR-QA complex question answering subset (Accuracy, %)
表 2. 各模型在 BAR-QA 复杂问答子集上的表现(准确率，%)

模型(Model)	跨章节问题准确率	复杂计算问题准确率
DocFinQA-style Retriever	41.3	38.5
FinSage-style RAG	55.8	52.1
LedgerLens	78.5	72.6

表 2 结果显示是，LedgerLens 的智能体框架展现了出色的泛化能力。其灵活的“规划–工具调用”模式使其能够自主处理需要多步推理的复杂问题，性能远超相对固化的 RAG 流程。这证明了本文的框架不仅是一个优秀的抽取器，更是一个强大的通用问答系统。

4.5.3. 消融实验：验证核心组件有效性

为了量化 LedgerLens 框架内部各核心组件的贡献，在 BAR-QA 数据集上进行了一系列消融实验。通过移除或替换关键模块来评估其对最终指标抽取性能(F1 分数)的影响，结果如表 3 所示。

Table 3. Ablation study results of LedgerLens on the BAR-QA metric extraction task
表 3. LedgerLens 在 BAR-QA 指标抽取任务上的消融实验结果

模型(Model)	F1 分数(%)	F1 下降幅度(%)
LedgerLens (完整版)	94.1	-
w/o Planner (使用原始查询)	90.5	−3.6
w/o Iteration (单轮检索)	88.2	−5.9
w/o Table Vectorization	91.8	−2.3

消融实验的结果清晰地揭示了各模块的价值：

移除规划智能体(w/o Planner)后，F1 分数下降了 3.6%。这表明，对于口语化或不标准的查询(如“净利”)，Planner 的规范化能力是保证后续检索准确性的第一道防线。

移除迭代机制(w/o Iteration)是影响最大的改动，导致 F1 分数骤降 5.9%，性能水平回落至与 FinSage-style RAG 相似的区间(87.8%)。这一结果表明，研究智能体的多轮迭代、自我修正能力正是 LedgerLens 超越传统 RAG 范式的根本原因。

使用常规表格向量化也导致了 2.3%的性能下降，这说明本文提出的“去数字化”嵌入策略能有效帮助模型聚焦于表格的语义结构，提升了表格检索的准确率。

4.5.4. 公开基准对标与效率分析

在证明了 LedgerLens 在核心任务上的优越性后，进一步通过次要实验来评估其泛化性和实用性。

Table 4. Officially reported performance of SOTA models on the FinanceBench dataset
表 4. SOTA 模型在 FinanceBench 数据集上的官方报告性能

模型(Model)	LLM 评估准确率(%)	人工评估准确率(%)
Islam *et al*.	-	19.00

续表

FinSage	49.66	57.05
DocFinQA-style Retriever (复现)	48.5	51.2
LedgerLens	63.4	66.8

如表 4 所示，LedgerLens 在 FinanceBench 上的表现同样超越了包括 FinSage 在内的所有已知方法，人工评估准确率达到了 66.8% [9] [10]。这证明了本文提出的 Agent 框架不仅在特定任务上表现优异，也具备强大的泛化能力，是金融问答领域一个全面领先的解决方案。

5. 结论与未来工作

尽管 LedgerLens 在特定任务上取得了显著成效，但其架构也为未来的研究开辟了若干富有前景的方向，旨在构建更为通用与强大的文档智能系统。

首先，在任务复杂度上，当前框架可从“单点信息定位”向“复杂分析推理”演进。本文工作的核心优势在于对单一、离散指标的精准抽取。然而，真实的金融分析场景常涉及需要整合多个数据点进行计算与比较的复杂任务，例如“计算某公司近三年的净利润复合年均增长率(CAGR)”。未来的研究应致力于提升规划智能体(Planner Agent)的任务分解(Task Decomposition)能力，使其能够将此类复杂分析查询自动拆解为一系列原子的信息抽取子任务。同时，需探索为研究智能体(Researcher Agent)引入状态记忆(Stateful Memory)与结果聚合(Result Aggregation)机制，使其能够缓存并综合多轮检索的中间结果，最终完成需要跨表格、跨章节的聚合计算与比较分析。这项工作将推动系统从一个高效的“信息抽取器”向一个具备初级分析能力的“推理引擎”演进。

其次，在系统鲁棒性上，应着力于构建面向异构文档的“端到端”解决方案。本研究的一个前提假设是上游 PDF 解析工具能够提供高质量的结构化文本。这一依赖性构成了系统的潜在瓶颈，因为在面对格式异常或版式复杂的表格时，解析错误难以避免，并将对下游任务的成败产生直接影响。为了构建一个真正端到端的解决方案，未来的工作应着力于融合多模态信息(Multimodal Information)以提升系统的容错性。具体而言，可以为智能体框架集成一个“视觉校验模块”(Visual Verification Module)。当智能体对检索到的文本块(尤其是表格)的结构完整性置信度较低时，可触发该模块，调用多模态大语言模型(Multimodal LLM)直接分析原始文档相应区域的视觉表征。通过比对视觉信息与文本信息，智能体可以自主识别并修正解析错误。该方向的研究将极大提升系统在真实、复杂和非理想文档环境下的适用性与可靠性。

参考文献

[1] Huang, Y., Chen, Y., Zhang, H., Li, K., Fang, M., Yang, L., Li, X., Shang, L., Xu, S., Hao, J., Shao, K. and Wang, J. (2025) Deep Research Agents: A Systematic Examination and Roadmap. 10.48550/arXiv.2506.18096.

[2] Zheng, Y., Fu, D., Hu, X., Cai, X., Ye, L., Lu, P. and Liu, P. (2025) Deep Researcher: Scaling Deep Research via Reinforcement Learning in Real-World Environments. ArXiv, abs/2504.03160.

[3] Sarmah, B., Mehta, D., Hall, B., Rao, R., Patel, S. and Pasquali, S. (2024) Hybridrag: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction. *Proceedings of the 5th ACM International Conference on AI in Finance*, Brooklyn, 14-17 November 2024, 608-616. https://doi.org/10.1145/3677052.3698671

[4] Wang, J., Ding, W. and Zhu, X. (2025) Financial Analysis: Intelligent Financial Data Analysis System Based on LLM-RAG. *Applied and Computational Engineering*, **145**, 182-189. https://doi.org/10.54254/2755-2721/2025.22221

[5] Setty, S., Jijo, K., Chung, E. and Vidra, N. (2024) Improving Retrieval for RAG based Question Answering Models on Financial Documents. ArXiv, abs/2404.07221.

[6] Kim, S., Song, H., Seo, H. and Kim, H. (2025) Optimizing Retrieval Strategies for Financial Question Answering

Documents in Retrieval-Augmented Generation Systems. ArXiv, abs/2503.15191.

[7] Lee, J. and Roh, M. (2024) Multi-Reranker: Maximizing Performance of Retrieval-Augmented Generation in the Finance RAG Challenge. ArXiv, abs/2411.16732.

[8] Reddy, V., Koncel-Kedziorski, R., Lai, V., Krumdick, M., Lovering, C. and Tanner, C. (2024) DocFinQA: A Long-Context Financial Reasoning Dataset. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 2: *Short Papers*), Bangkok, 11-16 August 2024, 445-458. https://doi.org/10.18653/v1/2024.acl-short.42

[9] Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N. and Vidgen, B. (2023) Finance Bench: A New Benchmark for Financial Question Answering. ArXiv, abs/2311.11944.

[10] Jimeno-Yepes, A., You, Y., Milczek, J., Laverde, S. and Li, R. (2024) Financial Report Chunking for Effective Retrieval Augmented Generation. ArXiv, abs/2402.05131.

[11] Jiang, C., Zhang, P., Ni, Y., Wang, X., Peng, H., Liu, S., *et al*. (2025) Multimodal Retrieval-Augmented Generation for Financial Documents: Image-Centric Analysis of Charts and Tables with Large Language Models. *The Visual Computer*, **41**, 7657-7670. https://doi.org/10.1007/s00371-025-03829-5

[12] Gondhalekar, C., Patel, U. and Yeh, F. (2025) MultiFinRAG: An Optimized Multimodal Retrieval-Augmented Generation (RAG) Framework for Financial Question Answering. ArXiv, abs/2506.20821.

[13] Lai, V.D., Krumdick, M., Lovering, C., Reddy, V., Schmidt, C.W. and Tanner, C. (2024) SEC-QA: A Systematic Evaluation Corpus for Financial QA. ArXiv, abs/2406.14394.

[14] Wang, X., Chi, J., Tai, Z., Kwok, T.S., Li, M., Li, Z., He, H., Hua, Y., Lu, P., Wang, S., Wu, Y., Huang, J., Tian, J. and Zhou, L. (2025) FinSage: A Multi-Aspect RAG System for Financial Filings Question Answering. ArXiv, abs/2504.14493.