金融 RAG 毕业论文（定稿草稿）

摘要

金融问答场景中的复杂查询往往涉及多年份、多实体与显式数值计算，传统单步检索 + 占位式生成在复杂问题上容易出现证据覆盖不足与算术错误。本文面向 FinDER（金融领域专家检索）数据集构建可复现的金融 RAG 系统，在无外部 LLM API 的约束下引入规则驱动的多步检索与显式计算器，并通过门控与系统化调参保证性能稳定。实验表明：检索器微调显著提升整体检索表现（full dev Recall@10: 0.3246→0.3772）；多步检索在复杂子集上保持不退化且 MRR 略有提升；计算器门控避免数值指标回退，为后续提升奠定稳定基线。所有实验产出与配置均可在 outputs/ 中追溯与复现。

关键词：金融问答；检索增强生成；多步检索；显式计算；误差分析；可复现

---

1 引言

金融领域的专业问题常涉及最新的市场数据、监管条款与财报细节，具有信息密度高、语义压缩强的特点。传统大型语言模型在金融问答中容易产生事实性错误或因知识滞后而偏离证据。为提高可靠性，检索增强生成（RAG）通过“检索—生成”机制使回答基于真实文档，但现有多数金融 RAG 仍以**单步检索**为主，面对复杂金融查询时表现不足。

复杂查询通常具备以下特征：
信息分散：所需证据散落于多个段落或多份文档；
- 结构化需求：涉及同比、差值、占比等显式计算；

语义压缩：包含金融缩写与术语（如“MS”“YoY”），需进行实体消歧与扩写。

- FinDER 数据集由金融领域专家构建，包含 5,703 个真实金融问答三元组（查询—证据—答案），其中大量查询简短含糊且依赖跨证据推理。该数据集凸显了“多步检索 + 结构化推理”的必要性。
- **本文思路**：引入规则驱动的多步检索推理机制，使系统能像分析员一样逐步补全证据链；同时引入显式计算模块，降低数值类问题的算术错误。系统在工程实现上坚持开源、可复现与可审计，确保实验结果可追踪与可复验。
- **本文贡献**：
- 1. 构建可复现的金融 RAG 工程框架，统一数据、检索、评测与实验编排；
- 2. 实现规则驱动的多步检索循环，结合 gap 检测、合并策略与门控机制；
- 3. 引入显式数值抽取与计算器模块，通过单位/年份校验与回退策略控制错误传播；
- 4. 在 full dev / complex dev / numeric dev 上进行系统化调参与对照评测，形成论文级结果表格与错误分析。
- ---

- 2 相关工作
-
- 与本文相关的研究可归纳为以下四类：
-
- 金融问答（Financial QA）
- 金融问答聚焦于财报、公告、研报等长文档中的事实性与数值型问题。该领域强调证据定位与实体对齐，且普遍存在缩写、口径不一致等问题。
-
- 检索增强生成（RAG）评测
- RAG 将检索与生成结合，近年来的评测工作关注检索器类型（稀疏/稠密/混合）、检索器微调对下游任务的传导，以及复杂问题上的检索覆盖能力。
-
- 多跳/多步检索
- 多跳检索通过多轮检索逐步补全证据链，常用于跨段或跨文档推理。其核心挑战是如何设计有效的“继续检索”判断与证据合并策略。
-
- 智能化 RAG（Agentic RAG）
- Agentic RAG 通过规划与工具调用扩展检索能力，但在成本、可控性与可复现性方面仍有挑战。本文选择规则化的多步检索与显式计算，以强调稳定性与可解释性。
-
- ---
-
- 3 方法
-
- 本文系统由四个核心模块构成：**查询理解**、**多步检索推理**、**证据整合与计算**、**答案生成**。系统采用流水线式数据流，模块间接口明确，便于复现与扩展。
-
- **图 1 多步检索循环流程（示意）**
-
- Query → Query Understanding → Step-1 Retrieval → Gap Detector
-   → (gap & gate) Refined Query Retrieval → Merge & Rank
-   → Calculator (optional) → Template Answer
-
- 3.1 查询理解
- 目标是将原始金融查询规范化，解决缩写、实体歧义与任务类型识别问题。可采用：
- 规则与词典扩展（如将“YOY”扩展为“同比”）；

金融领域 NER 与实体链接；

- 轻量语义解析（不开启外部 API）。

输出为规范化查询 q' 与结构化槽位（实体、指标、年份、计算类型），作为多步检索的输入。

3.2 多步检索推理

多步检索在第 t 轮使用当前查询 q_t 检索 top-k 证据，基于 gap 检测决定是否继续：

Gap Detector：判断是否缺失关键年份或比较对象；
- **Gate**：若 $\text{gap\_conf} < \tau$，终止后续检索；

Merge Strategy：跨步候选去重并按 maxscore 或 step1 extunderscore first 排序；
- **Stop Criteria**：达到 T 或连续无新增证据时停止。

关键超参数定义：
检索轮数 T（max_steps）
- 每轮检索 top-k（top_k_each_step）

最终截断 top-k_f（top_k_final）
- 门控阈值 τ（min_gap_conf）

Step6 最优配置：$T=2$，top_k_each_step=10，merge=maxscore，novelty_threshold=0.0，stop_no_new_steps=1，$\tau=0.3$。

3.3 证据整合与计算
该模块从证据中抽取数值、年份、单位与实体，并执行显式计算（YoY/差值/占比/倍数）。核心约束：
单位一致性校验；
- 年份对齐要求；

候选冲突时拒算并回退。
-
- **Step6 最优门控**：min_conf=0.2，allow_task_types=[]（当前版本以稳定性为先）。
-
- 3.4 答案生成
- 采用模板化生成：若计算器返回 status=ok 且通过门控，则输出结构化结果与解释；否则回退基线答案，并记录 fallback 原因以支持审计。
-
- ---
-
- 4 实验设置
-
- 数据集与划分
- 使用 FinDER 数据集，包含 5,703 个查询—证据—答案三元组。数据按 train/dev/test 划分，所有样本统一格式：
-
- { "qid": "...", "query": "...", "answer": "...", "evidences": [{"text": "..."}], "meta": {} }
-
- 子集定义
- **complex_dev**：满足任一条件即进入子集：多证据、查询含 ≥2 年份、含比较/变化关键词、或含数值+年份组合。

numeric_dev：查询或答案含数字/百分号/同比/差值/倍数关键词。
-
- 评价指标与口径
- 检索指标：Recall@k、MRR@k、evidence_hit@k

QA 指标：EM/F1（用于对照）

- 数值指标：Numeric-EM、RelErr、Coverage

不确定匹配比例：当证据缺少 doc_id/evidence_id 时使用文本匹配，并记录比例。

-
- 关键参数
- 检索器：BM25 + Dense + Hybrid（alpha=0.5）

多步检索（best）：max_steps=2，top_k_each_step=10，merge=maxscore

- 计算器门控（best）：min_conf=0.2，allow_task_types=[]

所有实验配置与结果均保存在 outputs/<run_id>/，可复现。

---

5 实验结果与分析

本节直接引用 Step6 输出的表格与指标（见 `docs/TABLE_MAIN.md`、`docs/TABLE_NUMERIC.md`），并对主要对照结果进行分析。

1) 检索效果（full dev / complex dev）

主结果表见：`docs/TABLE_MAIN.md`

关键对照（complex dev）：
baseline(post-ft) vs best multistep：

- - Recall@10：0.3909465 → 0.3909465（持平）
- - MRR@10：0.2960138 → 0.2960873（+0.00007）
-
- 检索器微调带来的整体提升（full dev）：
- pre-ft baseline → post-ft baseline：Recall@10 0.3246 → 0.3772（+0.0526）

2) 数值题表现（numeric dev）

数值表见：`docs/TABLE_NUMERIC.md`

关键对照（numeric dev）：
baseline(post-ft) vs best calc gate：

- - Numeric-EM：0.3838 → 0.3838（持平）
- - RelErr(mean)：683.3536 → 683.3536（持平）
- - Coverage：0.6266 → 0.6266（持平）
-
- 说明：当前版本计算器门控在 dev 上选择 allow_task_types=[]，以避免数值误差回退。因此 numeric 指标未下降，但尚未体现提升。该结果为“安全启用”基线，可在后续提升抽取/计算置信度后再重新开启任务类型。
-
- 3) 消融与案例
-

- 消融结果见 `docs/TABLE_ABLATION.md`

典型复杂查询案例见附录 B（3 个案例，包含多步检索每步 top-3 证据与 stop 原因）

-
- ---
-
- 6 错误分析与案例
-
- 1. 失败类型概览（Step6）
- **complex dev**：主要集中在 `no_gap` 与 `max_steps`，说明部分查询在第一步已覆盖核心证据，但 refine 能力仍有限；

numeric dev：由于计算器门控关闭（allow_task_types=[]），多数样本回退到 baseline，表现为 fallback 占比高。

-
- 2. 失败原因分析
- 1) **简称歧义与实体对齐不足**：金融缩写可能指向多家公司，导致检索命中无关证据。
- 2) **检索漂移**：refined query 若过于相似或偏离目标，会重复检索或引入噪声。
- 3) **证据冲突与口径不一**：不同段落可能存在统计口径差异（如合并口径与单体口径），需要更强的单位/实体对齐策略。
- 4) **计算器保守门控**：为避免算错，门控阈值偏保守，导致覆盖率受限。
-
- 3. 典型案例
- 典型复杂查询案例已整理在附录 B（`docs/CASE_STUDIES.md`），包含 3 个查询的多步检索轨迹与证据对比。
-
- ---
-
- 7 讨论
-
- 1) 多步检索的收益与限制
- 多步检索在 complex dev 上避免了退化，但整体提升有限，主要原因在于 gap 识别与 query 重写仍偏规则化，无法充分挖掘隐式实体关系与跨段落依赖。
-
- 2) 计算器的可控性与覆盖率
- 显式计算显著降低了算术错误的风险，但门控策略为了安全性关闭了多数计算任务，导致数值指标未提升。后续需通过更强的单位对齐与置信度校准逐步放宽门控。
-
- 3) 未来工作
- 引入更强的检索器与领域适配预训练；

提升实体消歧与 query rewriting 的准确性；

- 扩展计算器任务类型与多尺度单位转换；

结合轻量 agent 机制增强多步检索的决策质量。

-
- ---
-
- 8 结论

-
- 本文面向 FinDER 金融问答任务构建了可复现的金融 RAG 系统，在无外部 LLM API 的约束下引入多步检索与显式计算模块。实验表明：
-
- 检索器微调显著提升整体检索表现（full dev Recall@10: 0.3246 → 0.3772）；

多步检索在复杂子集上保持不退化，MRR 略有提升；
- 计算器门控避免数值指标回退，为后续优化奠定稳定基线。

未来工作将集中在提升 gap 识别与 query 重写能力、增强数值抽取与单位对齐鲁棒性，并探索更强的检索与推理模型。

附录 A 结果表格

表 A1 主结果（full dev 与 complex dev）

label	run_id	full_r10	full_mrr10	complex_r10	complex_mrr10
pre_ft_baseline	20260130_234540_ae7cdf_m01	0.3246	0.2030	0.3457	0.2330
post_ft_baseline	20260130_234540_ae7cdf_m02	0.3772	0.2601	0.3909	0.2960
post_ft_multistep_best	20260130_234540_ae7cdf_m03	0.3772	0.2601	0.3909	0.2961
post_ft_baseline_calc_best	20260130_234540_ae7cdf_m04	0.3772	0.2601	0.3909	0.2960
post_ft_multistep_calc_best	20260130_234540_ae7cdf_m05	0.3772	0.2601	0.3909	0.2961
post_ft_multistep_T1_calc_best	20260130_234540_ae7cdf_m06	0.3772	0.2601	0.3909	0.2960

表 A2 数值题结果（numeric dev）

label	run_id	num_em	num_rel	num_cov
pre_ft_baseline	20260130_234540_ae7cdf_m01	0.3791	2874.5248	0.6202
post_ft_baseline	20260130_23454	0.3838	683.3536	0.6266

	0_ae7cdf_m02			
post_ft_multistep_best	20260130_234540_ae7cdf_m03	-	-	-
post_ft_baseline_calc_best	20260130_234540_ae7cdf_m04	0.3838	683.3536	0.6266
post_ft_multistep_calc_best	20260130_234540_ae7cdf_m05	0.3838	683.3536	0.6266
post_ft_multistep_T1_calc_best	20260130_234540_ae7cdf_m06	0.3838	683.3536	0.6266

表 A3 消融结果

label	run_id	full_r10	full_mrr10	complex_r10	complex_mrr10
post_ft_multistep_T1_calc_best	20260130_234540_ae7cdf_m06	0.3772	0.2601	0.3909	0.2960

附录 B 典型复杂查询案例

5.4 典型复杂查询案例（3 个）

案例 1（qid=8c8c8c34）
Query：Hasbro (HAS) 2023 one-time charges impact on operating profitability vs historical trends and cap allocation implications.
- Gold Answer（摘要）：In 2023, Hasbro’s operating result turned from a profit in prior years (407.7 million in 2022 and 763.3 million in 2021) to an operating loss of 1,538.8 million…

Step0 Top3：008beea7_e0_c0，8c8c8c34_e0_c2，f8aec91a_e0_c1
- Step1 Top3：008beea7_e0_c0，f8aec91a_e0_c1，8c8c8c34_e0_c2

gap/stop：MISSING_ENTITY / MAX_STEPS，final_topk_size=10
- 分析：该问题包含对比关系与年份信息，多步检索识别到 gap，但 refined query 与原查询高度相似，导致新增证据有限。

案例 2（qid=52e25ec7）
Query：Impact on net investing cash flows from EUC sale cash inflow offsets vs acquisition outflows, AVGO.
- Gold Answer（摘要）：The $3,485 million inflow from the sale of the EUC business helped to partially offset the significantly higher cash expenditures related to acquisitions…

Step0 Top3：506e7d1e_e0_c0，52e25ec7_e0_c0，e4661352_e0_c3
- Step1 Top3：506e7d1e_e0_c0，52e25ec7_e0_c0，1c47856d_e0_c1

gap/stop：MISSING_ENTITY / MAX_STEPS，final_topk_size=10
- 分析：问题涉及“出售现金流入 vs 并购现金流出”的对比，多步检索能够维持证据覆盖但未显著扩展证据范围。

案例 3（qid=ed746c33）

Query：Cash flow & cap alloc implications of IRM's ASC 842 storage rev rec vs other lines.

- Gold Answer（摘要）：For its Global Data Center Business, Iron Mountain recognizes storage revenues under ASC 842⋯

Step0 Top3：ed746c33_e0_c0，2a8785e8_e0_c15，a68b8600_e0_c5

- gap/stop：NO_GAP / NO_GAP，final_topk_size=10

分析：该类问题实体明确、语义集中，单步检索即可覆盖核心证据，多步检索不引入额外噪声。

-
- 参考文献

- [1] 参考文献占位。

-