

研 究 计 划

面向上市公司公告风险识别的金融事件知识图谱构建与证据 增强推理研究

申 请 人	数字经济专业学生（脱敏）
目 标 院 系	浙江大学计算机科学与技术学院
目 标 导 师	陈华钧 教授
日 期	2026 年 8 月

（一）立项依据

1.1 研究背景与问题提出

上市公司公告集中记录股权质押、股权冻结、重大资产损失、破产清算和高管变动等关键事件，是理解企业经营变化与风险传播的重要公开文本。相较于行情和财务指标，公告中的事件信息具有长文本、跨句分布、多事件共存、时间更新频繁和专业表达不统一等特征。传统人工阅读能够保留上下文，但难以稳定处理大规模文档；仅依赖关键词或单句分类，又容易遗漏分散在不同段落中的事件论元及其时间关系。

文档级金融事件抽取为上述问题提供了结构化入口。DCFEE、Doc2EDAG 和 DocFEE 等研究已将中文金融公告转化为事件类型及其论元集合^[5-7]。其中，DocFEE 收录 19,044 篇公告、9 类事件和 38 类论元，平均文档长度超过 2,200 个中文字符，能够较真实地反映长距离论元依赖^[7]。然而，从公告抽取事件并不等于形成可用于风险分析的可靠知识：同一主体可能跨公告出现，事件状态会随时间更新，自动抽取结果也可能与原文冲突。若不保存来源片段、公告时间、抽取置信度和实体对齐记录，图谱中的三元组很难接受复核。

知识图谱能够连接企业、股东、金融机构、事件、时间和金额等要素，但专业领域构图还面临 schema 演化、实体粒度和原文证据对齐等问题。Schema-adaptable KGC 讨论了新类型不断出现时的动态 schema 适配^[3]；K-ON 指出知识图谱以实体为基本单元，而大语言模型以 token 预测，两者存在粒度不匹配^[1]；KAG 则通过知识图谱、原始片段互索引和混合推理提高专业知识服务的可解释性^[2]。这些工作说明，金融事件图谱的价值不在于“把文本变成图”，而在于能否把结构化事件、时间变化和可追溯原文组织成一条可检查的证据链。

因此，本研究拟以公开中文上市公司公告和开放金融事件数据集为对象，依次研究文档级事件抽取、事件知识图谱构建与证据增强风险识别。研究不把“预测股价”作为直接目标，也不预设复杂模型必然优于透明基线，而是重点回答：事件能否被稳定抽取，跨公告知识能否保持时间一致，风险判断能否回到原文证据，以及模型在不确定时能否拒绝给出过强结论。

1.2 国内外研究现状与评述

1.2.1 文档级金融事件抽取

DCFEE 较早提出面向中文金融文档的自动标注与事件抽取框架，突破了单句事件抽取对人工标注数据的依赖^[5]。Doc2EDAG 将文档级事件抽取表示为有向无环图生成任务，更好地处理多事件和论元跨句分布^[6]。DocFEE 进一步以真实长公告扩展数据规模，并通过人工核验评估半自动标注质量^[7]。这些研究为本计划提供了公开数据与可复现基线，但仍存在三类问题：不同公告中同一实体的对齐尚未闭合；预定义事件类型难以覆盖新增风险表述；抽取指标高并不意味着下游风险结论可靠。

1.2.2 动态知识图谱构建

通用文档级关系抽取研究表明，跨句推理和多实体关系是构图中的持续难点^[8]。Schema-adaptable KGC 将 schema 横向扩展、纵向扩展和混合扩展纳入统一测试，说明静态 schema 在动态领域中的局限^[3]。CodeKGC 则利用代码语言模型的结构先验和 schema-aware 提示进行生成式构图^[4]。对金融公告而言，事件类型、论元角色和主体关系具有较强规则性，但业务口径和披露文本会不断变化。因此，本研究将固定核心事件 schema，同时保留“候选新类型—人工确认—版本更新”的扩展路径，不把一次抽取结果直接写入权威图谱。

1.2.3 时间推理与风险识别

金融事件不是静态事实。股权质押可能发生补充质押、解除质押或冻结，企业风险状态也会随公告更新。RE-Net 以自回归结构预测时间知识图谱中的未来交互^[11]；TPAR 同时处理时间图谱插值与外推，并强调路径推理的可解释性与抗噪性^[9]。但高风险场景下，强制模型对所有样本给出判断会放大错误。Selective TKGR 引入置信度估计和拒答机制，使模型能够对低置信度预测保持沉默^[10]。本研究据此把“可选择性输出”作为风险识别的重要评价，而不是只追求覆盖率。

1.2.4 知识增强与证据可追溯

K-ON 从实体级预测角度探索知识图谱与大语言模型融合^[1]，KAG 则将知识图谱和原始文本片段双向索引，结合逻辑形式与语义推理用于专业问答^[2]。金融文本还包含数值、表格和多步计算，FinQA 显示金融文档理解需要同时处理文本、表格与显式推理程序^[12]。这些工作共同提示：风险分析系统应保存“结构化事件—图路径—原文片段—数值口径”的对应关系，并对来源冲突和证据缺失进行显式标记。

表 1 代表性研究及本计划的推进空间

文献	研究对象与方法	可直接借鉴	尚待推进
Yang 等 ^[5]	中文金融文档；自动标注与事件抽取	金融事件 schema 与远程监督思路	长文本、多事件与证据追溯
Zheng 等 ^[6]	文档级多事件；Doc2EDAG	多事件与跨句论元组合	跨公告实体对齐和时间更新
Chen 等 ^[7]	19,044 篇真实长公告；DocFEE	公开数据、长距离依赖和人工核验	从抽取走向图谱和下游风险任务
Ye 等 ^[3]	动态 schema 构图	schema 扩展与条件解码	金融事件类型的版本治理
Liang 等 ^[2]	KG+原文片段+混合推理	双索引与证据增强	金融公告中的来源冲突和数值口径
Hou 等 ^[10]	时间图谱选择性推理	低置信度拒答	将拒答与原文证据完整度结合

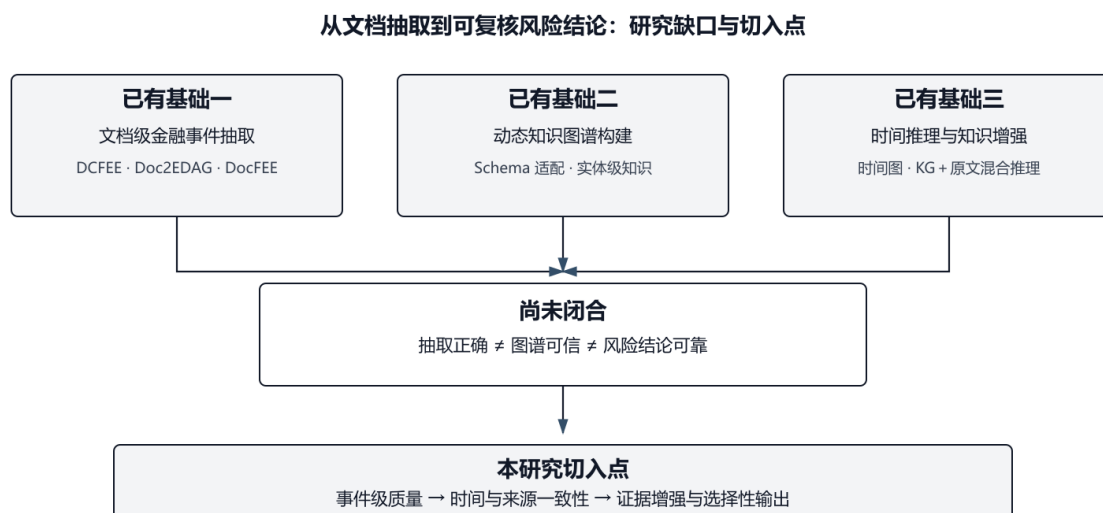


图 1 文献基础、未闭合问题与本研究切入点

1.3 拟解决的关键研究问题

关键研究问题一：长公告中的金融事件如何稳定抽取并保持证据可追溯？需要比较序列标注、文档级图生成和 schema-aware 生成式抽取，评价事件类型、论元角色和事件级完整性，并为每条结构化结果保留公告标识、原文跨度、发布时间与抽取置信度。

关键研究问题二：跨公告事件如何形成具有时间一致性的知识图谱？ 需要处理企业简称、股东与机构名称对齐，区分事件发生、披露、更新和解除时间，并在 schema 变化或来源冲突时保留版本，不以覆盖率为理由合并不确定实体。

关键研究问题三：结构化事件与原文证据能否共同提高风险识别的可靠性？ 研究将比较表格特征基线、事件图特征、图路径+原文证据三种输入，重点检查跨时间和跨企业外推、概率校准、证据覆盖率以及低置信度拒答，而不是把模型贡献直接解释为因果作用。

1.4 研究价值

理论上，本研究把金融事件抽取、动态知识组织和风险识别连接为统一证据链，讨论事件事实、时间状态与风险判断之间的层级边界。方法上，研究把事件级抽取指标、图谱一致性和选择性风险预测纳入连续评价，避免“抽取准确率高”直接替代下游可靠性。应用上，研究输出可追溯的事件记录和条件性风险提示，为金融文本分析提供可复核原型，但不替代投资、审计或监管决策。

在数据治理层面，本研究将来源管理嵌入模型流程。每条事件记录不仅包含事件类型和论元，还保留公告标识、发布时间、原文跨度、抽取模型与 schema 版本。这样可以区分“原文已经披露的事实”“模型依据文本形成的结构化记录”和“下游模型给出的风险判断”。当事件被补充、解除或更正时，研究保留前后版本及其关系，不覆盖历史记录。该设计有助于定位误差发生在文本抽取、实体对齐、时间组织还是风险模型，而不是在最终指标异常后才追查数据来源。

在评价层面，本研究强调模块间误差传递。事件类型判断正确但关键金额或时间缺失，仍可能导致图谱状态错误；图谱局部连接完整但实体被错误合并，也可能使企业风险特征被污染。因此，除分别报告抽取、构图和预测指标外，还将选取端到端样例，逐步检查公告原文、事件记录、图路径和模型输出是否一致。若下游性能提升主要来自少数高频企业或时间泄漏，则该提升不能作为方法贡献。

1.5 研究目标与逻辑闭环

本研究设置三个递进目标。第一，形成可复现的中文金融公告事件抽取基线，明确长文档、多事件、跨句论元和少样本类型的主要误差。第二，构建保留实体候选、时间状态、来源片段和版本信息的最小事件知识图谱，并给出可量化的图谱质量报告。

第三，在严格时间切分和企业留出条件下，检验图结构与原文证据能否改善风险识别的校准和外推表现，并让系统对证据不足样本主动拒答。

三个目标构成“数据事实—知识组织—条件判断”的逻辑闭环。目标一没有达到事件级完整性要求时，只保留高置信样本进入图谱；目标二未解决来源和时间冲突时，不进入风险识别；目标三若不能在替代切分中复现，则保留透明基线并报告负结果。通过这些升级条件，研究的最终价值不依赖某个复杂模型取得最高分，而取决于每一级结论是否具有与其强度相匹配的证据。

（二）研究内容与研究方案

2.1 总体目标与研究边界

本研究拟构建“文档级金融事件抽取—时间事件知识图谱—证据增强风险识别”的最小可行框架，并在公开数据上评价其准确性、稳定性和可追溯性。

研究设置四条边界。第一，公告抽取结果是机器生成记录，未经人工核验不视为权威事实。第二，风险标签用于方法验证，不等同于企业真实违约概率或投资建议。第三，SHAP、图特征贡献和注意力权重仅解释给定模型的预测行为，不作为因果证据。第四，若数据授权、标签质量或时间覆盖不足，研究缩小为公开数据复现与方法比较，不形成现实业务结论。

2.2 研究内容一：金融事件 schema 与文档级抽取

以 DocFEE 的 9 类事件和 38 类论元为公开起点，建立事件类型、主体、客体、金额、比例、发生时间、披露时间和状态字段的数据字典。对新增类型采用候选池管理：模型提出候选，人工判断是否与已有类型同义或确属新增，再更新 schema 版本。

建立三类基线：（1）句级或段落级序列标注，用于观察局部模型的上限与遗漏；（2）Doc2EDAG 类文档级抽取，用于处理多事件和跨句论元；（3）schema-aware 生成式抽取，用于检验新增类型和结构约束。评价包括事件类型宏平均 F1、论元 F1、事件级完整率、跨句论元召回率和无效结构率，并进行公告长度、事件数量与少样本类型分组分析。

2.3 研究内容二：时间事件知识图谱构建

将抽取结果组织为“企业—事件—参与主体—时间—数值—来源”图结构。企业实体以证券代码为主键，名称作为时间相关属性；自然人、机构和子公司实体使用标准

化名称与上下文约束进行候选对齐。对低置信度对齐保留独立节点和候选链接，避免错误合并并在图中扩散。

事件节点同时记录发生时间、披露时间和状态更新时间。对同一事件的补充公告、进展公告和解除公告建立 `updates` 或 `resolves` 关系，并保存来源文档与原文跨度。图谱质量通过实体对齐准确率、时间冲突率、重复事件率、来源覆盖率和抽样人工复核一致性评价。

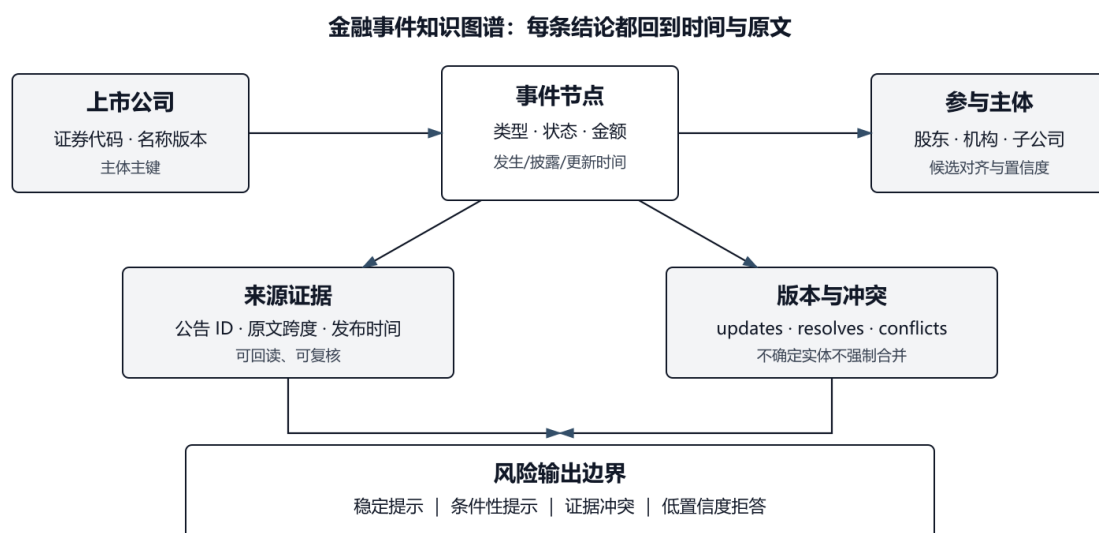


图 2 金融事件图谱中的实体、时间、来源与结论边界

2.4 研究内容三：证据增强风险识别

在不构造投资建议的前提下，选择一个可公开复现的演示标签，例如未来一段时间内是否出现同类风险事件、监管关注或财务异常。设置三组输入：第一组为传统财务与公告统计特征；第二组增加事件类型、频次、时间间隔和图邻域特征；第三组进一步增加与预测相关的图路径和原文证据片段。

模型从逻辑回归和 LightGBM 开始，再根据数据规模比较图表示或时间图基线。数据划分采用严格时间切分，并补充企业留出验证，避免同一企业的未来信息泄露到训练集。评价除 AUC、宏平均 F1 和 Brier 分数外，还包括概率校准、跨时间性能衰减、企业留出性能、证据覆盖率和选择性预测风险。只有在加入图谱或证据后，性能与校准同时改善且在替代切分中可复现，才认为其提供了稳定增益。

2.5 技术路线

研究按“问题界定—数据治理—事件抽取—时序图谱—风险识别—综合验证”七个阶段展开，并归入四个递进研究层。层间采用单一主轴，层内保留由抽取误差和时间冲突分析驱动的两条研究迭代；研究任务、评价指标与预期成果的对应关系如图 3 所示。

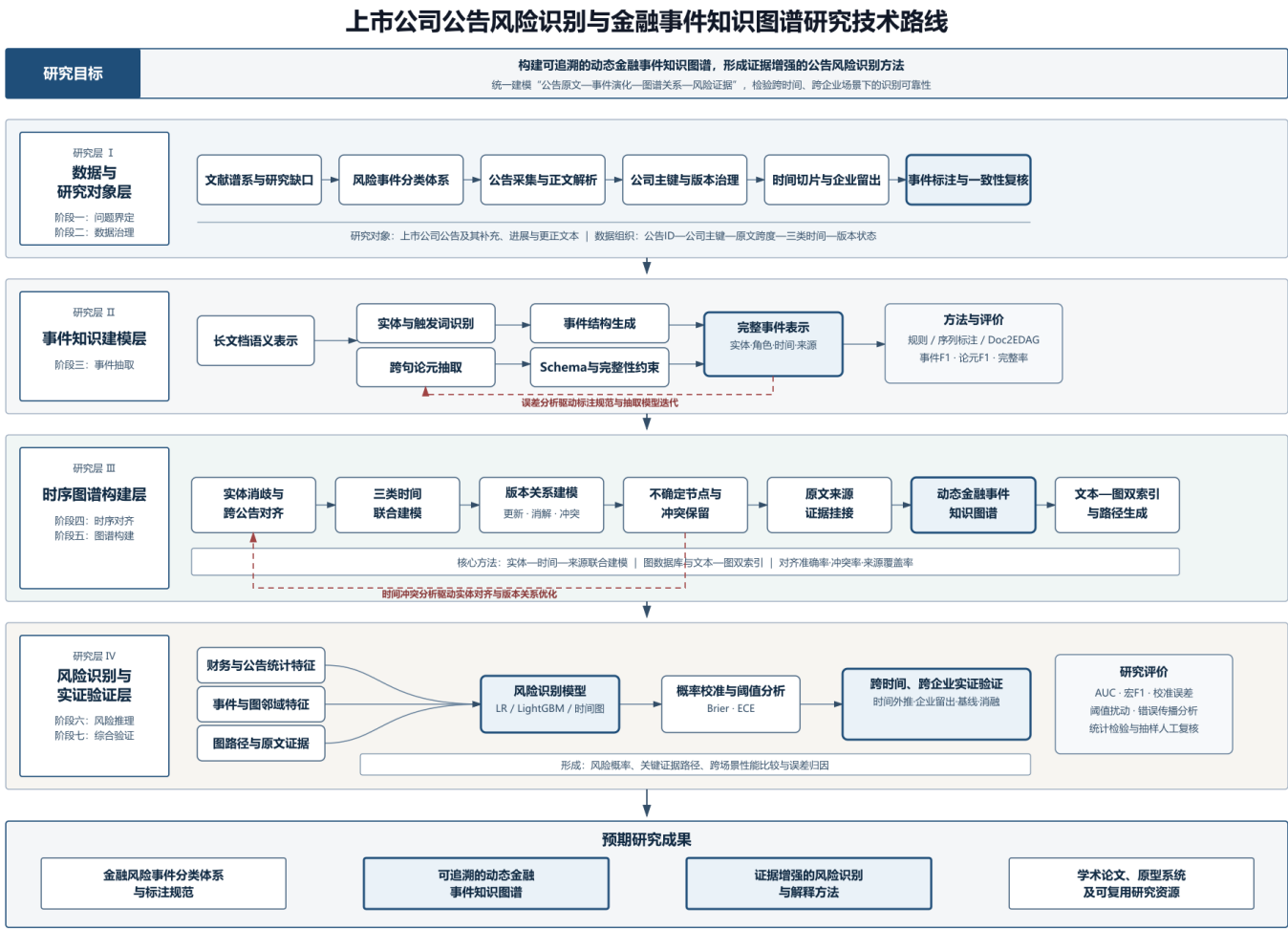


图 3 上市公司公告风险识别与金融事件知识图谱研究技术路线

图 3 不仅给出研究推进顺序，还把每一阶段的研究对象、技术任务、评价指标和预期成果对应起来。其中，两条局部迭代分别由事件抽取误差和时间冲突分析驱动：前者用于修订标注规范并优化抽取模型，后者用于改进实体对齐与版本关系建模。最终形成金融风险事件分类体系与标注规范、可追溯的动态金融事件知识图谱、证据增强的风险识别与解释方法，以及论文、原型系统和可复用研究资源。

表 2 研究问题、方法与可检查产出的对应关系

研究问题	主要方法	关键评价	可检查产出
长公告事件能否稳定抽取	序列标注、Doc2EDAG、schema-aware 生成	事件/论元 F1、完整率、无效结构率	数据字典、抽取基线、误差表
跨公告知识能否时间一致	实体对齐、事件去重、版本与来源管理	对齐准确率、时间冲突率、来源覆盖率	图谱 schema、样例子图、质量报告
风险判断能否证据增强	LR、LightGBM、图/时间表示、双索引检索	校准、外推、证据覆盖、拒答风险	风险基线、消融、证据链案例

2.6 验证方案与稳健性评价

上游抽取采用官方训练/验证/测试划分，并增加按文档长度和事件数量分层的误差分析。图谱构建对实体对齐、事件合并和时间状态分别抽样复核，记录错误是否会传递到下游任务。风险识别采用时间外推为主、随机切分仅作对照；同一企业跨集合泄漏和未来公告泄漏均作为阻断性错误。

稳健性评价覆盖数据、schema、模型和时间四个维度。数据维度改变标注置信阈值和样本纳入规则；schema 维度比较核心类型与扩展类型；模型维度比较透明基线、树模型与图表示；时间维度比较不同训练窗口和测试区间。对于风险输出，报告覆盖率—风险曲线：随着拒答比例提高，保留预测的错误率是否下降。若模型无法识别自己的不确定性，则不把其标记为“可用于风险提示”。

表 3 主要失败情形与替代路径

失败情形	触发条件	替代路径	结论边界
生成式抽取结构不稳定	无效结构率或漏论元显著高于基线	回退受约束解码或 Doc2EDAG 基线	不以模型规模替代抽取质量
实体对齐错误扩散	抽样复核发现高频错误合并	提高阈值，保留候选链接与人工复核队列	图统计只基于高置信子图

失败情形	触发条件	替代路径	结论边界
风险标签噪声过大	标签与公告事件定义不一致	改为同类事件复现或排序演示任务	不输出企业风险概率
图特征无稳定增益	时间切分或企业留出中增益消失	保留透明特征基线并分析失败	不宣称知识图谱提高预测
证据不足或相互冲突	图路径无法回到原文或来源冲突	输出“证据不足/冲突”并拒答	不生成单一确定结论

2.7 预期创新与可能贡献

- （1）把金融事件抽取的事件级正确性、知识图谱的时间一致性和风险识别的选择性可靠度组织为一条连续评价链。
- （2）为每个事件节点保留“公告—原文跨度—抽取置信度—schema 版本”，使图谱推理结果能够回到原始证据。
- （3）把拒答纳入金融风险识别评价，以覆盖率—风险曲线区分稳定提示、条件性提示与证据不足，避免对所有企业强制输出结论。

（三）研究基础与进度安排

3.1 已完成的课程与方法训练

本人已接受数字经济、计量经济学、产业经济学和信息经济学等课程训练，现有材料显示 GPA 为 4.23/5，专业排名 5/178，计量经济学等核心课程成绩在 96 分以上。这些课程为企业行为理解、变量构造与实证结果阅读提供了基础。

在企业数字化项目研修中，本人接触 2012—2023 年 A 股 CSMAR 数据，项目样本为 41,687 条，涉及面板数据清洗、双向固定效应、调节效应、IV-2SLS、稳健性与异质性分析。该经历可迁移到公告主体对齐、企业时间切分、风险特征构造与结果稳健性评价。由于个人具体分工仍需在正式提交前复核，本计划仅将其作为项目研修和方法训练，不表述为独立主持成果。

在机器学习项目中，本人使用 1,000 条合成结构化样本比较 Ridge、随机森林、SVR、XGBoost 和 LightGBM，并使用 SHAP 解释模型，现有记录中 LightGBM 测试集 R^2 为 0.8695。该经历说明本人具备多模型比较和可解释分析的入门基础，但合成数据结果不代表真实行业结论。本研究将补充充分评价、概率校准、时间外推和解释稳定性训练。

3.2 已完成工作、可迁移能力与明确缺口

表 4 已完成工作与待补能力

已完成工作	可迁移到本研究的能力	当前缺口	前期补齐方式
A 股面板数据清洗与变量构造研修	企业一年份数据治理、缺失与异常检查	非结构化公告解析	复现 DocFEE 数据读取和长文档切分
固定效应、IV、稳健性与异质性训练	切分意识、替代设定与边界解释	事件抽取评价	复现 DCFEE/Doc2EDAG 基线与误差分析
LightGBM 与 SHAP 项目	透明基线、多模型比较、局部解释	知识图谱与图推理	学习 RDF/ 属性图、Neo4j/NetworkX 和实体对齐
数据治理兴趣	来源、口径和结论边界意识	KG-RAG 与证据链	实现图节点到原文跨度的双向索引

3.3 待完成工作与二十四周安排

本研究先完成公开数据复现和最小图谱，再逐步增加证据增强与风险任务。若真实公告下载、导师数据或算力未落实，公开 DocFEE 数据上的抽取、构图和评价工作仍可独立完成，不影响前十二周产出。

表 5 二十四周研究进度与阶段验收

阶段	时间	主要任务	阶段成果	验收标准
文献与复现准备	第 1—4 周	阅读导师代表工作与金融事件抽取文献；补 NLP/KG 基础	文献矩阵、术语表、复现环境	研究问题与每个方法选择均有文献依据
数据与抽取基线	第 5—8 周	读取 DocFEE；建立 schema 和基线；完成误差分层	数据字典、基线结果、误差表	无数据泄漏；事件与论元指标可复现
图谱最小闭环	第 9—12 周	实体标准化、事件去重、时间与来源建模	图谱 schema、样例子图、质量报告	每个事件可回到公告和原文跨度
风险识别基线	第 13—16 周	构造演示标签；LR/LightGBM；时间和企业留出	基线指标、校准图、泄漏审查	时间切分严格；结论不写成投资建议
证据增强与消融	第 17—20 周	加入图特征、路径与原文证据；拒答评价	消融表、覆盖率—风险曲线、案例	增益跨切分复现；低置信预测可拒答
稳健性与成稿	第 21—24 周	替代 schema/ 阈值 / 窗口；统一图表与口径	研究报告、代码说明、汇报材料	问题—方法—证据—边界闭合

3.4 可行性与入组后起步方案

本研究具有三方面可行性。其一，DocFEE、DCFEE 和 Doc2EDAG 提供公开数据、代码或清晰的定义，能够从复现开始而非依赖未获授权的内部数据。其二，学生已有 A 股数据和机器学习基础，可以承担数据清洗、透明基线与稳健性分析，同时把 NLP/KG 列为明确补课任务。其三，研究将复杂目标拆成抽取、构图和风险识别三个可验收模块；即使下游任务不成立，上游数据与图谱质量评价仍能形成真实负结果。

入组后前 30 天完成核心文献矩阵、DocFEE 数据字典和一个透明事件抽取基线；第 31—60 天完成文档级模型复现与分层误差分析；第 61—90 天形成带原文证据的最小事件图谱和图谱质量报告。风险识别与知识增强只在上游质量达到导师确认的阈值后进入，避免将未经核验的抽取噪声包装为应用成果。

3.5 与导师研究主线的适配及推进边界

陈华钧教授公开研究方向包括人工智能、知识图谱、自然语言处理和 AI for Science，近年代表工作涉及动态 schema 构图、知识图谱与大语言模型融合以及专业领域知识增强^[1-4]。本计划以金融公告为专业领域载体，将上述方法问题具体化为事件抽取、时间一致性和可追溯推理；同时利用学生已有的 A 股数据、稳健性评价和可解释机器学习训练，形成可执行的前 90 天任务。

本选题是基于公开主线形成的导师讨论模板，不代表导师已确定该课题。正式研究对象、数据、模型深度和成果形式应在交流后共同确认；如导师更关注通用知识图谱构建，本计划可保留事件抽取和证据链主线，弱化金融风险任务；如数据条件不足，则限定为公开数据复现与图谱质量研究。

3.6 预期成果

预期形成一套金融事件 schema 与数据字典、一个可复现的文档级事件抽取基线、一个保留时间与原文来源的最小事件知识图谱、一组时间外推和选择性风险识别实验，以及包含失败结果和结论边界的研究报告。代码、配置、数据处理说明和不含受限数据的样例可作为复现材料保存。

参考文献

- [1] Guo L, Zhang Y, Bo Z, et al. K-ON: Stacking Knowledge on the Head Layer of Large Language Model[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(11): 11745-11753. DOI: 10.1609/aaai.v39i11.33278.
- [2] Liang L, Sun M, Gui Z, et al. KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation[EB/OL]. arXiv:2409.13731, 2024.
- [3] Ye H, Gui H, Xu X, et al. Schema-adaptable Knowledge Graph Construction[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.
- [4] Bi Z, Chen J, Jiang Y, et al. CodeKGC: Code Language Model for Generative Knowledge Graph Construction[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024.
- [5] Yang H, Chen Y, Liu K, et al. DCFEE: A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data[C]//Proceedings of ACL 2018, System Demonstrations. 2018: 50-55. DOI: 10.18653/v1/P18-4009.
- [6] Zheng S, Cao W, Xu W, et al. Doc2EDAG: An End-to-End Document-level Framework for Chinese Financial Event Extraction[C]//Proceedings of EMNLP-IJCNLP 2019. 2019: 337-346. DOI: 10.18653/v1/D19-1032.

- [7] Chen Y, Zhou T, Li S, et al. A dataset for document level Chinese financial event extraction[J]. Scientific Data, 2025, 12: 772. DOI: 10.1038/s41597-025-05083-9.
- [8] Yao Y, Ye D, Li P, et al. DocRED: A Large-Scale Document-Level Relation Extraction Dataset[C]//Proceedings of ACL 2019. 2019: 764-777. DOI: 10.18653/v1/P19-1074.
- [9] Chen K, Wang Y, Li Y, et al. A Unified Temporal Knowledge Graph Reasoning Model Towards Interpolation and Extrapolation[C]//Proceedings of ACL 2024. 2024: 117-132. DOI: 10.18653/v1/2024.acl-long.8.
- [10] Hou Z, Jin X, Li Z, et al. Selective Temporal Knowledge Graph Reasoning[C]//Proceedings of LREC-COLING 2024. 2024: 14555-14566.
- [11] Jin W, Qu M, Jin X, et al. Recurrent Event Network: Autoregressive Structure Inference over Temporal Knowledge Graphs[C]//Proceedings of EMNLP 2020. 2020: 6669-6683. DOI: 10.18653/v1/2020.emnlp-main.541.
- [12] Chen Z, Chen W, Smiley C, et al. FinQA: A Dataset of Numerical Reasoning over Financial Data[C]//Proceedings of EMNLP 2021. 2021: 3697-3711. DOI: 10.18653/v1/2021.emnlp-main.300.