ChemSafe-KG

从知识图谱到数据洞察——

化工安全事故分析的工程实践与效果验证

数据库技术及应用 · 期末项目技术总结报告

翟彝凡 余亮阳 赵乐毅

清华大学化学工程系

2026 年 6 月

1. 研究动机与背景

化工安全事故具有小概率、大后果的特征。单起事故可能导致数十人伤亡和巨大经济损失。应急管理部每月发布事故简报，但这些信息以非结构化文本形式存在，难以系统查询和模式归纳。一线安全人员面对具体问题时，无法高效地从数百份简报中获取答案。

大型语言模型（LLM）能够理解自然语言并从文本中提取信息，但存在一个关键问题：模型回答问题时引用的信息无法追溯来源。我们预实验发现，用 DeepSeek 直接回答 20 道化工安全问题，平均每题引用 5.95 个不在事故数据库中的实体。这些实体可能来自模型的训练数据，有的可能正确，但用户无法区分哪些陈述有据可查、哪些是模型自行补全的。在化工安全领域，不可审计的回答存在风险。

知识图谱（KG）以结构化方式存储实体及其关系，天然支持可追溯的推理。近年来，LLM 与 KG 的结合（Graph RAG）被视为控制 LLM 输出的有效手段。但这一方法在中文化工安全场景中的实际效果尚缺乏系统性验证。

相关工作

化工安全事故的分析方法经历了三个阶段。最早是传统统计分析——关文玲、蒋军成（2007）统计了 2001-2006 年化工火灾爆炸事故的频次和设备类型分布。这种方法能揭示宏观趋势，但无法追踪因果链。2018 年前后出现了事理图谱方向的研究，用知识图谱建模事件间的因果演化关系，但建图过程依赖人工编写规则，成本高、扩展难。2024 年微软开源 GraphRAG 框架后，LLM 自动构建知识图谱成为可能。GRAG-ProSafe QAS（2026）在钢铁行业 198 份英文报告上验证了这一思路，但未做与纯 LLM 的严格对比实验，也未涉及中文场景。

2. 问题定义与挑战

本项目的核心问题是：在中文化工安全事故场景下，用 LLM 自动构建知识图谱并约束 LLM 问答，能否实际提升回答的可控性和可追溯性？具体分解为三个子问题：（1）能否用 LLM 从非结构化的中文事故报告中自动构建可用的知识图谱？（2）建成之后，图谱对 LLM 问答的实际约束效果如何？（3）如果效果不理想，根因是什么？如何改进？

项目面临四个主要挑战。第一，中文化工事故报告长句多、术语密、格式不统一，传统基于规则的信息抽取方法难以覆盖。第二，LLM 抽取质量的评估不能仅凭个别案例，需要系统性的对照实验设计。第三，三个数据源的格式、粒度、可信度完全不同，整合需

要大量清洗和对齐工作。第四，数据量达到 1,335 份报告，人工标注不可行，必须依赖全自动化流水线。

3. 解决方法

我们设计并实现了端到端的五层系统架构，覆盖从数据采集到问答交互的完整链路。

3.1 Prompt Chain 知识抽取

知识抽取是系统的基石。采用 DeepSeek v4-flash 模型，设计了 Prompt Chain 策略，包含 8 条迭代规则：事件原子化（每个实体不超过 15 字）、人员操作分离（操作工误开阀门必须拆为 Equipment 和 Abnormal Condition 两个实体）、Equipment 和 Material 优先识别、5 种实体类型和 3 种关系类型、Few-shot 示例（丙烯腈储罐爆炸的完整因果链）、JSON 输出格式约束、实体类型白名单过滤。每条规则来自对失败抽取案例的归纳，Prompt 前后迭代了十几个版本。v0.7.1 引入 EntityNormalizer 对抽取结果做规范化。

容错方面实现了 JSON 三级恢复机制：空响应用更高温度（0.7）配合精简 Prompt 重试；Markdown 代码块自动去除；极端情况下用正则表达式强行提取 event_chain 字段。200 线程并发调用 LLM API，Phase 1 全并发收集结果，Phase 2 串行写入数据库以避免事务冲突。整体抽取成功率 99% 以上。

3.2 双存储架构

系统采用 Neo4j 5.26.25 图数据库和 SQLite 关系数据库的双存储架构。Neo4j 存储因果链——A 导致 B 这类关系天然适配有向图，变长路径查询一行 MERGE 即可完成。对 Equipment、Material、Accident 三种实体类型设置 UNIQUE 约束，防止重复节点。v0.7 版本新增 Accident 聚合节点，每起事故的所有相关实体通过 belongs_to 关系归组。SQLite 存储结构化事故记录（1,174 条去重后）、化学品物性（72 种）和天气记录（108 条），适合 SQL 统计分析。DataLinker 模块负责双向属性同步和一致性校验。

3.3 三层实体匹配与 Graph RAG

问答检索采用三层实体匹配策略。L1 精确匹配（jieba 分词后直接比对 KG 节点名，权重 x3），L2 关键词命中（分词与实体名交叉对比计数，权重 x1），L3 嵌入语义匹配（sentence-transformers 模型，470MB 多语言版本，带自适应阈值和实体名清洗，权重 x3）。融合公式为 score = L1x3 + L2x1 + L3x3，对 Top 30 实体额外查询 Neo4j 出边，有出边的节点加 2 分路径奖励。

Graph RAG 的核心是约束生成。检索层先通过三层匹配定位实体，再执行 Cypher 双向变长路径查询（max depth 4），经过去重和子路径过滤后，将格式化上下文传给 LLM。约束 Prompt 包含三条硬规则：严格按检索路径回答不得推测；路径不足以支撑回答时回复无法回答；每条陈述标注来源 [路径 N]。这一设计保证了答案的每条陈述都可追溯到图谱中的具体因果路径。

4. 数据流水线

数据流水线分为六个阶段，由 rebuild_all.py 脚本统一编排。

第零步：清空 Neo4j 和 SQLite 的既有数据。

第一步：数据采集。从 mem.gov.cn 的 95 个月度汇编页面用 BeautifulSoup 解析出 1,261 份事故报告（每份平均 150 字），从微信公众号化工安全教育服务平台获取 74 篇事故分析文章。

第二步：LLM 知识抽取。逐文件读取文本，通过 Prompt Chain 调用 DeepSeek v4-flash（temperature=0.5, max_tokens=16384），200 线程并发抽取，输出结构化 JSON（event_chain、root_cause、consequence）。

第三步：双库写入。Neo4j 端通过 batch_create_triples 方法以事务批量 MERGE 节点和 CREATE 关系，每条事务处理 30 条三元组，将网络往返从 O(n) 降至 O(1)。SQLite 端通过 SQLAlchemy ORM 逐条 INSERT。

第四步：数据充实。化学品方面，从事故记录统计频次，建立中英文名映射，通过 PubChem API 批量查询物性（CAS 号、分子量、闪点、爆炸极限、毒性分类），将化学品从 29 种扩展至 72 种。天气方面，从事故标题提取地名并映射经纬度，通过 Open-Meteo Archive API 获取历史天气，获得 108 条匹配记录。事故类型通过 classify_types.py 做一次性关键词分类。

第五步：事故去重。通过标题相似度（difflib.SequenceMatcher ratio > 0.95）识别重复记录，排除汇编文章系列后，标记 405 条重复（26%），去重后数据集 1,174 条。去重映射保存为 dedup_mapping.json。

第六步：验证与导出。运行健康检查确认节点数、关系数、事故数符合预期，导出 accidents.csv、chemical_properties.csv、weather_records.csv 和 causal_triples.csv 四份数据集。

关于数据代表性：本项目的数据来源均为公开渠道（政府月度简报和微信公众号），这决定了一个重要局限——我们的数据集代表的是被记录和被公开的事故，而不一定等同

于全部化工安全事故。未被公开报道的小型事故、企业内部未上报的未遂事件、以及因敏感原因未披露的重大事故，均不在此数据集中。此外，mem 月度简报的收录标准在不同年份可能有所变化，2020 年代事故数量的下降可能部分反映了收录标准或公开政策的变化，而非纯粹的安全形势改善。这一问题在后续研究中需要通过多源交叉验证来评估。

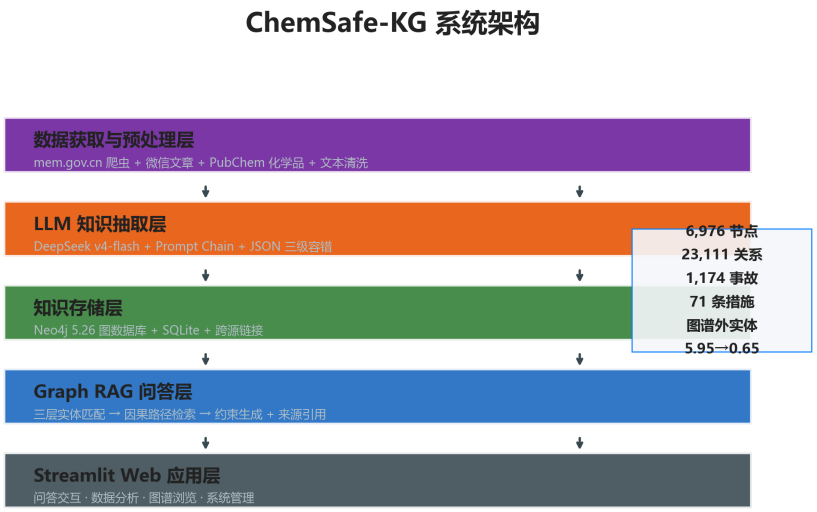

图 1: ChemSafe-KG 五层系统架构

5. 实验分析

5.1 实验设计

为检验知识图谱对 LLM 问答的约束效果，设计了三组对照实验。三组使用同一模型 DeepSeek v4-flash，唯一变量是检索方式。关键词 RAG 组：jieba 分词后从 SQLite 做文本检索 Top 8，LLM 基于检索结果回答。Graph RAG 组：三层实体匹配定位 KG 节点，Cypher 因果路径检索（max 4 跳），约束生成并标注来源。纯 LLM 组：不提供任何外部数据，仅凭模型参数知识回答。实验规模为 20 道题 x 7 种因果模式 x 4 维评估指标。每道题预设标准答案节点列表（人工标注 3-8 个关键实体）。

5.2 评估指标

图内约束率：回答中图谱外实体数为零的题目占比。来源可追溯率：答案中包含 [路径 N] 标注的题目占比。诚实拒答率：答案明确表示无法回答或说明信息局限的题目占比。平均图谱外实体数：每题回答中不在 Neo4j 实体白名单内的实体平均数。需要说明的

是，图谱外实体不等于事实错误——纯 LLM 引用的某些实体可能来自训练数据且正确——这一指标衡量的是可追溯性而非正确率。

5.3 实验结果

Graph RAG 的图内约束率为 70%（纯 LLM 5%），平均图谱外实体数从 5.95 降至 0.65，诚实拒答率从 0 提升至 55%。约束机制本身是有效的。

然而，实际使用体验并不理想。三个问题突出：一是检索速度随节点规模增长而下降（嵌入模型加载 6 秒，6,000 实体的相似度计算开销大）；二是无关因果链大量混入回答，碎片化的节点无法呈现清晰的事故因果结构；三是统计分析（SQL GROUP BY + 关键词匹配）产出的 14 条洞察反而比图谱 QA 更有实际价值。这些发现表明：在简报式数据（平均 150 字）上，图谱的投入产出比不合理。约束有效，但内容质量受限于源材料。

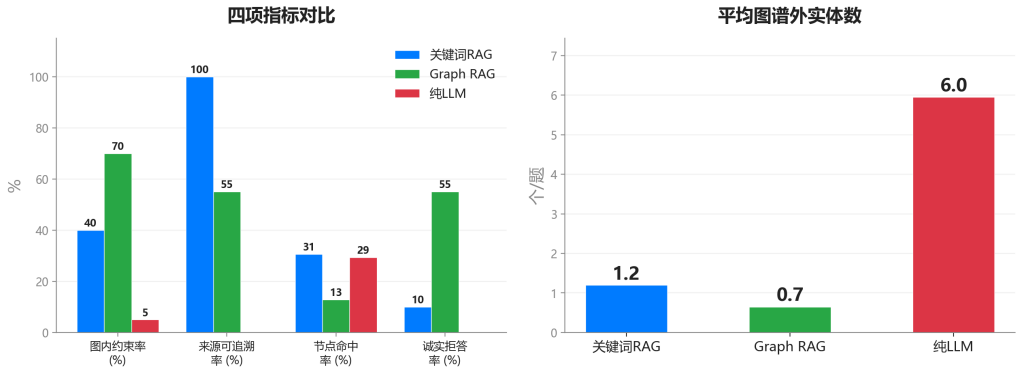

图 2: 三组对照实验结果对比

5.4 根因分析

效果不理想的根因有四个。第一，数据源天花板：月度简报 150 字的信息量不足以支撑深层因果推理，LLM 抽取的因果链平均深度仅 1-2 跳。第二，节点碎片化：同义实体未合并（形成爆炸性混合气体和形成爆炸性混合物是两个独立节点），导致检索时无关因果链大量混入。第三，响应速度退化：节点越多检索越慢，规模增长与可用性下降同步。第四，统计分析更直接：事故类型、频次、地域分布等最有价值的洞察用 SQL 即可得到，不需要知识图谱。

5.5 深层分析：2020 年代事故数量下降的多因素解读

数据洞察发现 2020 年代事故数量（95 起）相比 2010 年代（398 起）大幅下降 76%。这一趋势背后可能存在多重因素的交织。第一，安全生产政策效应：2016 年起实施的安全

生产综合治理三年行动计划，以及后续的化工企业搬迁改造政策，可能确实降低了事故发生率。第二，数据收录变化：月度简报的收录标准在疫情前后可能有所调整——2020-2022 年期间，应急管理部门的工作重心部分转移至疫情防控，事故报告的完整性和及时性可能受到影响。第三，产业结构调整：十三五期间大量化工企业从东部沿海向内陆迁移或关停并转，化工产能的变化也会影响事故的基数。第四，统计周期效应：2020 年代仅统计了前 5 年数据，若按年化计算，年均约 19 起，仍低于 2010 年代的年均 40 起。要区分这些因素的各自贡献，需要政策文本分析、产业结构数据和更长时序的事故记录进行交叉验证。这一分析本身也说明了数据科学的边界：统计能揭示趋势，但因果解释需要更多维度的信息。

6. 数据产品

项目交付了两个主要数据产品。

（1）Streamlit Web 应用。包含五个页面：系统概览（实时 Neo4j 统计）、因果推理问答（自然语言输入、三层匹配、Graph RAG 约束生成、来源引用）、多维数据分析（6 个标签页含 14 个统计洞察问答和 10 余种交互式 Plotly 图表）、知识图谱浏览（度中心性种子查询、barnesHut 物理布局、500+ 节点交互探索）、系统管理（配置检查）。

（2）开源数据集。以 CC BY-NC 4.0 许可在 GitHub Release v0.7.1 发布，包含四份 CSV 文件：accidents.csv（1,174 条去重事故记录）、chemical_properties.csv（72 种危化品物性）、weather_records.csv（108 条历史天气）、causal_triples.csv（Neo4j 导出的因果关系三元组）。数据集仓库地址：https://github.com/zyfxsxb45/ChemSafe-KG

6.1 数据洞察：14 条统计发现

项目最具实际价值的产出之一是从 1,174 条事故记录中系统性地挖掘出 14 条数据洞察。这些洞察全部基于 SQL 统计和关键词匹配，不依赖知识图谱，却提供了丰富的化工安全态势信息。

事故类型分布：爆炸 530 起（45.1%）、中毒窒息 210 起（17.9%）、其他 311 起（26.5%）、火灾 64 起、泄漏 58 起。爆炸和中毒窒息合计占 63%。

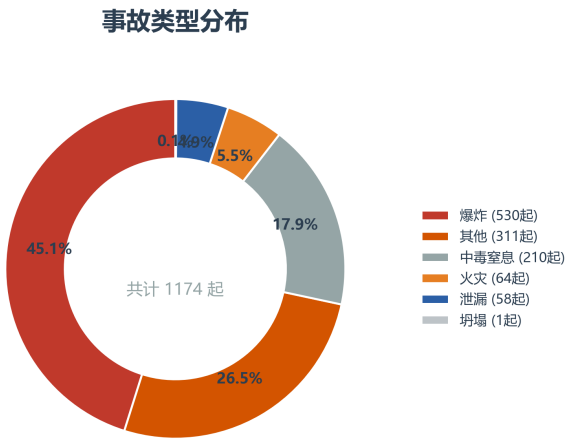

图 3: 事故类型分布

年代趋势：2010 年代为绝对峰值（398 起），2020 年代降至 95 起，降幅 76%。

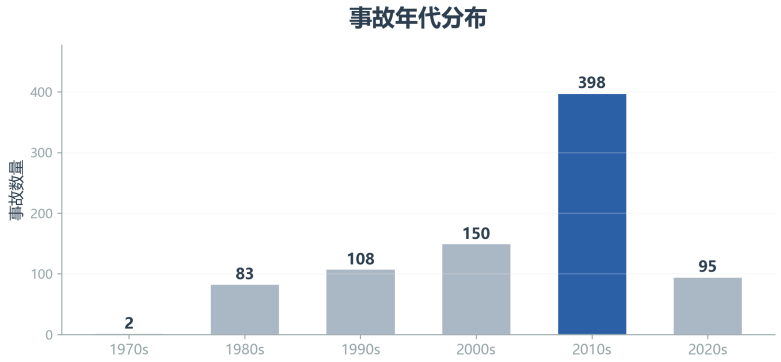

图 4: 事故年代分布

地域分布：山东 90 起、江苏 55 起、河北 50 起居前三，集中在东部沿海化工产业密集省份。

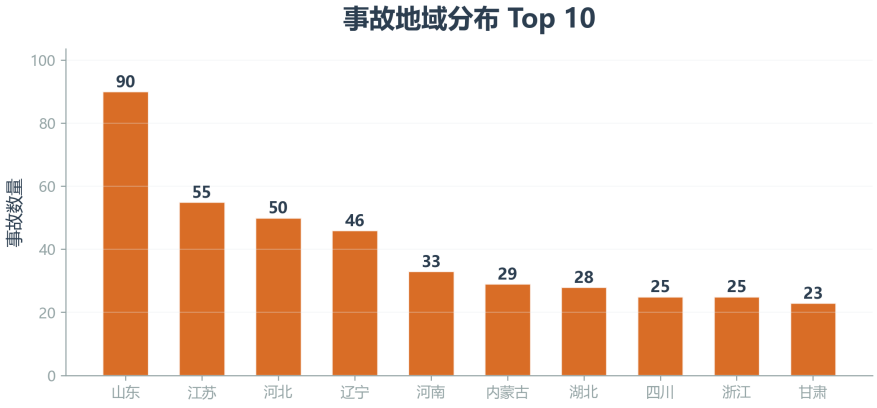

图 5: 事故地域分布 Top 10

化学品频次：氮气 35 起、硫化氢 30 起、甲醇 28 起。值得注意的是氮气本身无毒，但高浓度窒息是有限空间作业的典型危险。

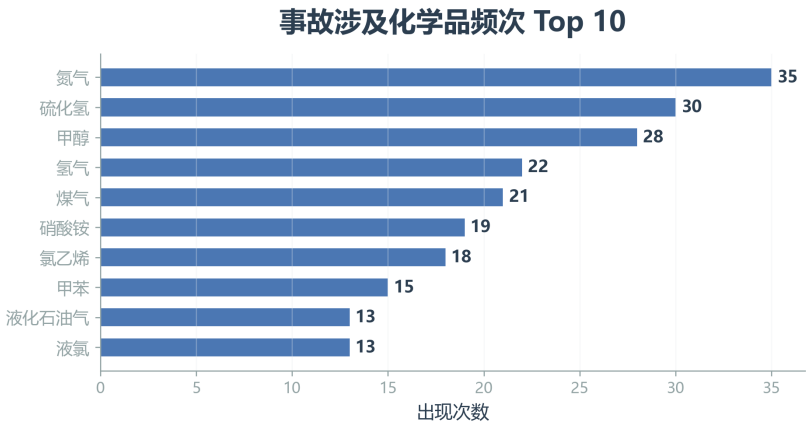

图 6: 事故涉及化学品频次 Top 10

设备频次：反应釜 40 起，是第二名管道（13 起）的三倍以上，反映了高温高压反应装置在化工事故中的核心地位。

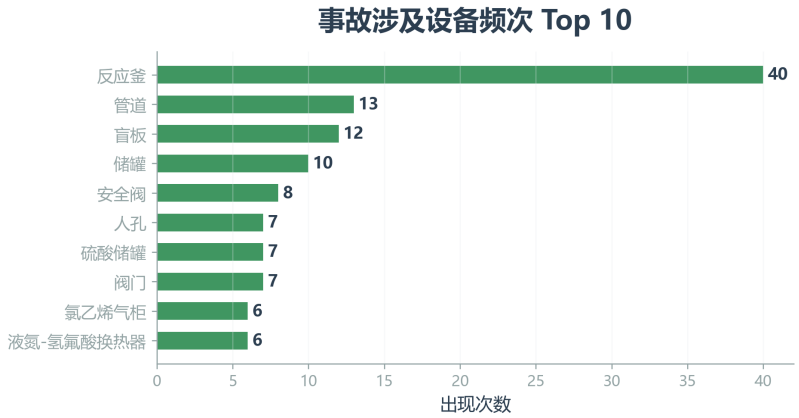

图 7: 事故涉及设备频次 Top 10

根因分布：违规操作占 31%（未办理动火票、未通风检测、未佩戴防护装备为高频词），设备故障占 27%（腐蚀、老化、泄漏为主要表现），两项合计近六成。

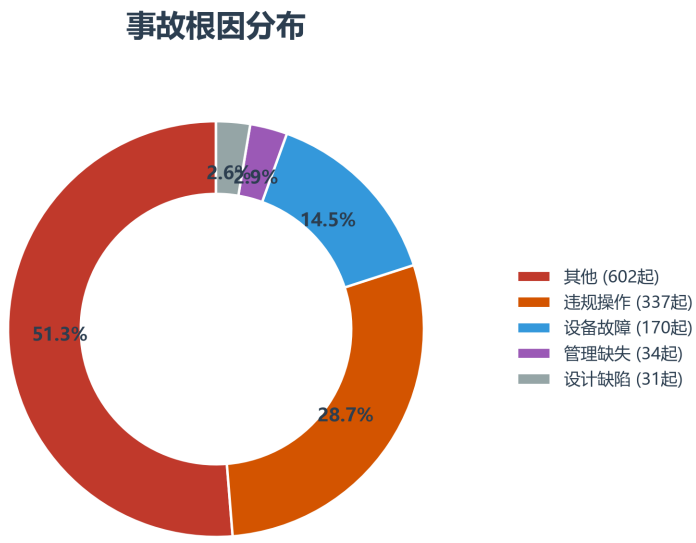

图 8: 事故根因分布

盲目施救：121 起事故涉及盲目施救或施救不当，占 10%，是有限空间作业中一人遇险、多人施救、全部遇难这一典型模式的数据印证。

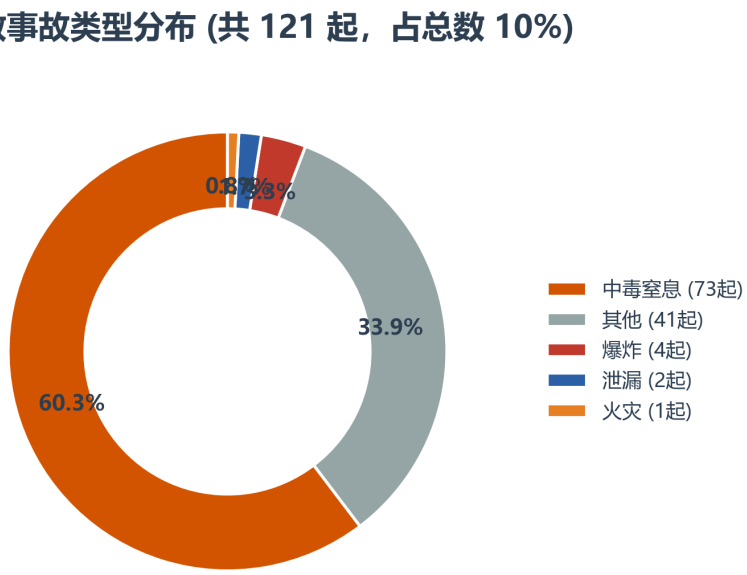

图 9: 盲目施救事故类型分布

闪点与事故频率：闪点低于 0 摄氏度的化学品平均事故频率为 10.7 起每品种，闪点高于 23 摄氏度的仅 3.8 起每品种，低闪点化学品的事故频率是后者的 2.8 倍。

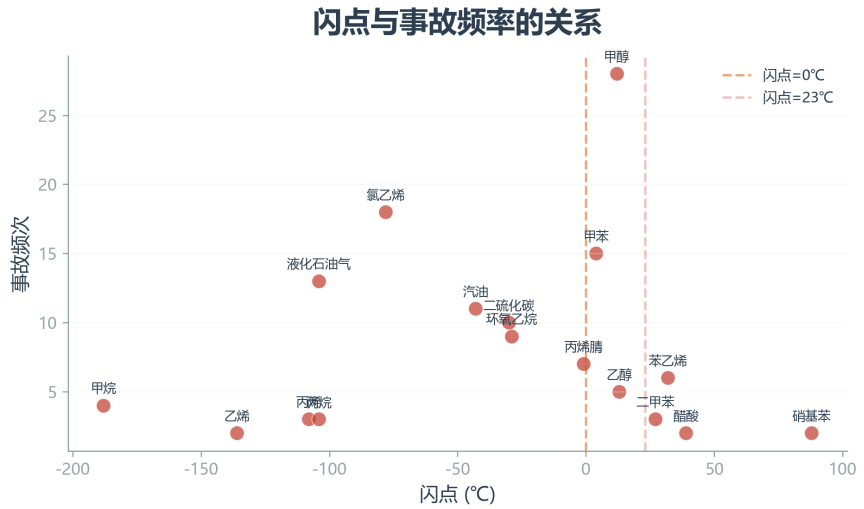

图 10: 闪点与事故频率的关系

季节性模式：一个反直觉的发现是，冬季爆炸占比（62.4%）略高于夏季（60.3%）。虽然夏季事故总数更多，但那是所有事故类型都增加的结果，而非爆炸事故独有。这一发现纠正了夏季爆炸更多的直观判断。

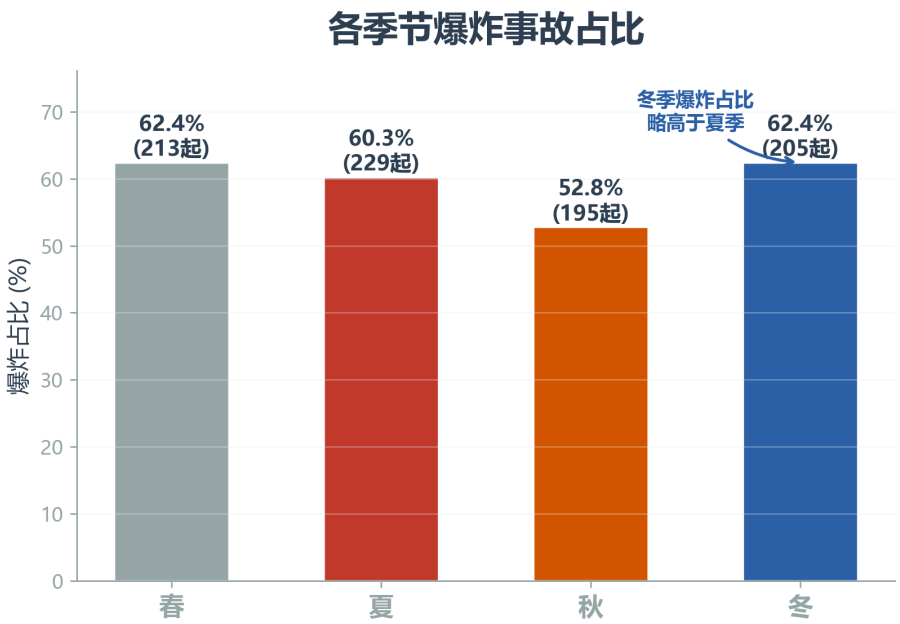

图 11: 各季节爆炸事故占比

月度分布：4 月事故最多（127 起），9 月最少（81 起），春季安全检查前后的事故报告集中可能是原因之一。

图 12: 月度事故分布

数据源对比：mem 简报 848 条（去重后）覆盖了基础事故信息，但 326 条微信文章在化学品和设备信息丰富度上显著更高，是 Mitigation 节点的主要来源。

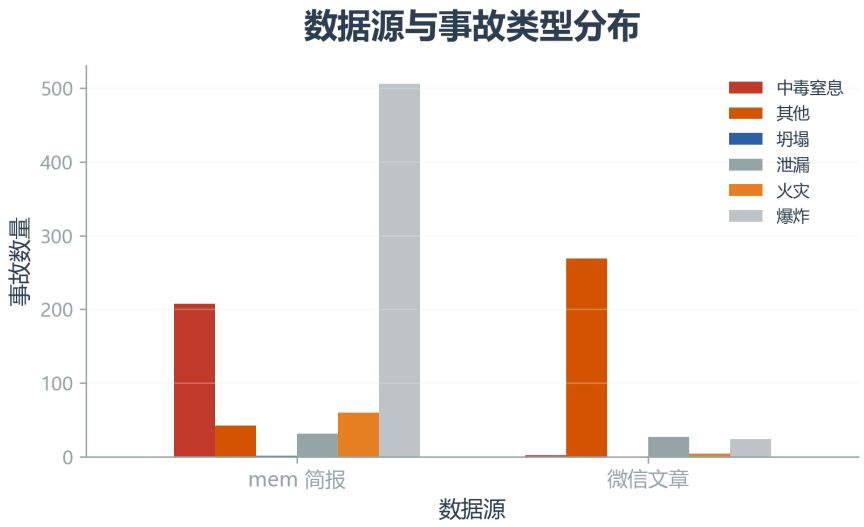

图 13: 数据源与事故类型分布

这些洞察的产出过程也带来了一个重要认知：最有价值的分析结果往往来自简单的统计方法加上对领域知识的理解，而非复杂的技术架构。这对于后续研究的方法选择具有指导意义——在数据源质量有限的条件下，深入理解数据比堆砌技术更重要。

7. 经验与教训

7.1 技术阻碍与攻克

Prompt 迭代是最耗时的工作。初版只有 5 条规则，LLM 输出的实体名经常是二十多个字的长句。通过逐批分析失败案例，归纳出事件原子化规则（不超过 15 字），并逐步加入人员操作分离、Equipment/Material 优先等约束。每条规则背后是几十条失败样本的

归纳。这一过程教给我们的是：LLM 辅助开发中，迭代速度比初始正确率重要——你第一天写不出完美的 Prompt，但你可以第一天跑起来看它哪里错。

JSON 解析容错是另一个反复调试的模块。LLM 经常输出不完整的 JSON（被 max_tokens 截断）、带 Markdown 代码块包裹的 JSON、或包含尾逗号和注释的非标准 JSON。三级恢复机制（空响应重试、代码块去除、正则强提）将解析成功率从约 85% 提升至 99% 以上。max_tokens 从 4096 逐步调至 16384 后，截断问题基本解决。

图谱可视化遭遇了框架限制。streamlit-agraph 的 Config 硬编码了 physics 参数结构，不允许自定义 barnesHut 的引力系数和弹簧长度。从随机节点布局到 hierarchical 再到 barnesHut，前后尝试了四个方案才找到可用的效果。最终使用度中心性种子查询加连通分量遍历，限 500 节点以内渲染。

7.2 核心经验

第一，数据质量是天花板。150 字的月度简报决定了因果链的深度上限，无论怎样优化 Prompt 都无法超越源材料的信息量。选择合适的数据源比花时间调算法更重要。

第二，对照实验是检验假设的唯一方法。我们的初始假设是 KG 能显著提升化工安全问答效果，实验证明了约束机制有效但实际效果不理想。这个结论推翻了初始假设，但方法论本身——同模型、同问题、唯一变量——是值得的且可迁移到其他项目的。

第三，测试自动化是最大的工程债。十几版 Prompt 迭代全部靠手动跑例子，早期缺乏自动化测试，目前端到端回归测试仍不足。如果重来，第一件事就是建立人工标注的黄金测试集和自动化评估流程。

8. 项目总结

ChemSafe-KG 是一个完整的化工安全事故知识图谱构建与效果验证系统，同时也是一次从知识图谱方法到数据洞察发现的研究旅程。我们从一个明确的假设出发——知识图谱能提升 LLM 在化工安全问答中的可控性——构建了覆盖数据采集、LLM 抽取、双存储、Graph RAG 检索和 Web 前端的五层系统，处理了 1,335 份事故报告，建成了 6,976 节点、23,111 关系的知识图谱，并在去重后形成了 1,174 条事故的高质量数据集。

通过三组对照实验（关键词 RAG、Graph RAG、纯 LLM，同模型唯一变量），我们验证了约束机制的有效性：图谱外实体从 5.95 降至 0.65，图内约束率从 5% 提升至 70%，诚实拒答率从 0 提升至 55%。但也发现了关键局限：在简报式数据（150 字/篇）上，图谱的碎片化、检索噪声和响应退化使得投入产出比不合理。统计分析产出的 14 条洞察反而更有实际价值。

项目的核心贡献不是证明了 KG 行或证明了 KG 不行，而是通过系统性实验厘清了 KG+LLM 方法在化工安全领域的适用边界，同时通过 14 条统计洞察展示了：即使在没有复杂技术架构的条件下，严谨的数据分析也能产出有价值的领域认知。约束机制有效，但需要配合高质量的长文本数据源（如 CSB 调查报告）、实体消歧、以及按事故类型构建的专用小图才能真正发挥 KG 在跨事故因果模式挖掘、失效树分析和风险传导路径建模方面的独特优势。而统计分析作为基线方法，在简报式数据上反而更为直接、可靠和具有解释力。我们锁定了让 KG 有效的必要条件，也验证了简单方法在合适场景下的价值。

数据集以 CC BY-NC 4.0 许可开源发布，代码在 GitHub 托管。项目虽为课程作业，但其数据工程实践、实验方法论和边界条件分析对后续的化工安全知识图谱研究具有参考价值。需要说明的是，数据集代表的是被公开记录的事故，不等同于全部化工安全事故的总体现状，使用时需考虑这一覆盖偏差。