处理 RAGged 事件：RAG 增强事件知识库构建与基于证明助手的推理

Stergios Chatzikyriakidis[0000−0001−8016−9743]

Computational Linguistics and Language Technology lab
Department of Philology
University of Crete
stergios.chatzikyriakidis@uoc.gr

Abstract. 从叙述文本中提取历史事件的结构化计算表示在手动构建时仍然计算成本昂贵。虽然 RDF/OWL 推理器支持基于图的推理，但它们仅限于一阶逻辑的片段，无法进行更深层次的时间和语义分析。本文通过使用多种大型语言模型（GPT-4、Claude、Llama 3.2）开发自动提取历史事件的模型，并采用三种增强策略来解决这两个难题：纯基生成、知识图谱增强和检索增强生成（RAG）。我们使用修昔底德的历史文本进行了全面评估。我们的研究发现，增强策略优化了不同的性能维度，而不是提供通用的改进。在覆盖范围和历史广度方面，基础生成通过 Claude 和 GPT-4 提取全面事件，达到了最佳性能。然而，为了提高精确度，RAG 增强改善了坐标的准确性和元数据的完整性。模型架构从根本上决定了增强的敏感性：较大的模型在基线性能上表现出稳健性，伴随着增量的 RAG 改进，而 Llama 3.2 则从表现优异到彻底失败显示出极大的差异。我们随后开发了一个自动翻译管道，将提取的 RDF 表示转换为 Coq 证明助手规格，从而能够进行超越 RDF 能力的高阶推理，包括多步骤因果验证、使用公元前日期的时间算术以及关于历史因果关系的正式证明。Coq 形式化验证表明，RAG 发现的事件类型代表合法的特定领域语义结构，而不是本体论违规行为。该研究确立了优化增强策略选择需要依赖用例特定的评估标准，而非一刀切的方法，同时展示了如何将自动本体构建奠定在形式验证之上以用于计算人文学科应用。[1]

1 引言

本体在计算语言学和人工智能中发挥着关键作用，通过以结构化的方式编纂知识，从而实现复杂的信息检索和推理。这些形式化的知识表示使机器能够理解实体之间的语义关系，促进超越简单关键词匹配的复杂查询。DBpedia 本体 [19] ，具有 768 个类，排列在一个包含关系的层次结构中，并由超过 3,000 个属性描述，展示了此类结构化知识库在增强自然语言理解方面的丰富潜力。

尽管结构化知识提取很有用，但它面临两个基本挑战，这两个挑战激发了对检索增强生成（RAG）的广泛研究。首先，手动构建事件知识库在计算上仍然昂贵且劳动密集。其次，现有的 RDF/OWL 推理机制在根本上局限于一阶逻辑的片段，限制了它们处理复杂时态语义、多步骤因果关系以及历史分析关键的复杂分析查询的能力。

尽管 RAG 在各种自然语言处理任务中展示了成功，但其对于领域特定知识提取的有效性仍然是一个未解的经验性问题。当大型语言模型已经对记录详尽的领域拥有大量内部知识时，外部知识增强普遍提高性能的普遍假设可能不成立。此外，模型架构与增强策略之间的相互作用——特别是对于像历史事件提取这样的专业化任务，目前还缺乏系统的研究。

本文通过系统比较多种大语言模型架构和增强策略的自动历史事件知识库提取，对传统关于 RAG 收益的假设提出了全面的实证分析。我们研究了三种不同的方法：不使用外部知识的基础生成、知识图谱增强及完整的 RAG 增强，应用了三种最先进的模型：GPT-4o、Claude-3.5 和 Llama

[1] 代码和实验数据可以在 https://github.com/StergiosCha/RAGGed_events 获得。

3.2。我们的分析揭示了一些违反直觉的发现，这些发现从根本上重新塑造了对外部知识增强何时以及如何应用的理解。我们的研究揭示了一个“反校准原则”，其中增强的有效性与模型能力呈反比：较强的模型如 GPT-4o 和 Claude-3.5 通过纯基础生成实现了更优异的性能（分别提取出 36 和 39 个全面事件），而外部增强通常有害——引入幻觉，降低覆盖率，并破坏语义准确性。相反，较弱的模型如 Llama 3.2 需要外部支撑，但表现出对实现质量的极端敏感性，其性能从具有竞争力到因增强复杂性导致的灾难性失败不等。

这些研究结果表明，增强策略优化了性能的不同维度，而不是提供普遍的改进。基础生成在全面的历史覆盖和叙述丰富性上表现出色，而 RAG 增强则以牺牲事件量和语义一致性为代价，换取技术精度——提高了坐标准确性和元数据的完整性。这种权衡表明，最佳策略的选择必须由特定的使用案例需求驱动，而不是盲目应用增强方法。

为了应对通过我们的提取分析揭示的 RDF/OWL 系统的基本表达能力局限性，我们开发了一条自动化翻译管道，将提取的知识表述转化为 Coq 证明助手的形式规格。这种翻译使得在传统语义网框架中不可能实现的高阶推理能力成为可能，包括验证多步骤因果关系、处理历史日期的复杂时间算术，以及关于历史因果关系的形式证明。Coq 形式化还验证了一个重要的理论发现：当根据通用知识图评估时，RAG 发现的事件类型似乎是本体论违规，但实际上它们代表了适合严格数学证明的合法的特定领域语义结构。

我们的方法使用修昔底德的《伯罗奔尼撒战争史》中的历史文本作为系统评估的受控领域。这样的选择使我们能够在有据可查的历史事件中评估增强策略，同时揭示模型架构与领域特征如何相互作用以决定最佳知识整合方法。将自动提取与形式化验证相结合，为结合现代自然语言处理规模与数学严谨性的计算人文学应用奠定基础。这项工作的核心贡献在于展示增强策略的有效性基本上依赖于模型能力、领域特征和预期应用需求之间的相互作用，挑战了领域内认为更全面的外部知识必然导致更好表现的假设。

2 相关工作

检索增强生成（RAG）是一种将基于检索的系统和生成模型的优势结合的方法，前者从外部来源获取相关信息，后者将提取的信息合成为连贯且具有上下文意识的输出 [20] 。RAG 使大型语言模型（LLM）能够动态访问最新的和特定领域的知识，从而提高其事实准确性，减少幻觉，并支持更细致的推理 [13] 。

RAG 模型已经在各种任务中取得了最先进的成果，包括开放域问答、信息检索和知识支撑的文本生成 [18] 。通过利用外部存储（如维基百科或专门的知识库），RAG 架构在生成更加具体、多样化和事实支撑的响应方面优于纯参数模型 [16] 。典型的 RAG 工作流程涉及检索相关的文档或段落，将这些信息与输入提示相结合，然后使用 LLM 生成最终的输出 [20] 。

除了通用问答之外，RAG 技术已被成功应用于信息抽取（IE）任务，包括命名实体识别、关系抽取和事件抽取 [34] 。在这些情况下，RAG 通过在检索到的、具有上下文相关性的证据中进行标识来增强性能和可解释性。例如，在命名实体识别中，检索到的相似句子可以改善语义表示，而在事件论元抽取中，增强检索的模型有助于捕捉跨论元依赖关系并提高少样本学习性能 [21] 。

最近的研究专门探讨了用于事件参数提取的 RAG，这是一项涉及识别和分类文本中描述的事件参与者和属性的任务 [10] 。检索增强生成模型，如 R-GQA [11] 和 DRAGEAE [14] ，检索类似的 QA 对或示例来指导生成过程，与抽取式和标准生成基线相比，表现出显著的改进。这些方法在标注数据有限的情况下特别有效，因为它们可以利用相关上下文中的示例来指导提取 [25] 。

RAG 已用于时间事件关系提取，以解决文本中时间语义的固有歧义问题 [37] 。通过检索来自 LLMs 的相关知识并将其整合到提示模板中，基于 RAG 的方法在提取和推理事件时间关系方面实现了卓越的性能 [23] 。

另一个相关的研究方向是将 RAG 与结构化本体相结合，例如特定领域的知识图谱 [2] 。本体基础的 RAG（OG-RAG）方法基本上是构建领域文档的超图表示，并根据本体关系对事实性知识进行聚类 [7,24] 。

最近对 RAG 配置的比较研究表明，检索复杂性和性能之间的关系是非线性的 [9] 。我们的工作通过展示在领域特定知识提取中，特别是在文献丰富的历史领域中，适度的检索策略可以优于最小和最大的外部知识整合，为这一新兴理解做出了贡献 [26] 。这挑战了 RAG 文献中广泛存在的假设，即更全面的检索必然导致更好的性能，突出了在 RAG 系统设计中进行领域特定优化的重要性 [1] 。

最后，[6] 使用 Coq 形式化了自然语言的时态和体态语义，作为符号自然语言推理系统的一部分，实现了符号方法的最新成果。总体而言，Coq 已被证明是一种强大的工具，可以形式化自然语言语义 [8] 。

我们的提议与一些相关工作相似，但在主要方法上却是独特的。我们提供的方案有两个方面：a) 使用 RAG 进行自动事件知识库的提取，并系统地比较不同的 RAG 和非 RAG 配置，和 b) 将提取的事件翻译到 Coq 证明助理中，从而实现超越简单 RDF 语言 [15,28] 的高级时序和因果推理。

3 方法论

提出的方法实现了一个两阶段的流程，用于将非结构化的历史叙事转化为正式的知识表示。第一阶段使用多种自然语言处理技术进行语义事件抽取，而第二阶段执行逻辑翻译以实现自动推理和形式验证：

- 阶段 1
  - 事件边界检测和分割
  - 带有语义角色分配的代理识别
  - 地理实体解析与坐标映射
  - 时间表达式规范化
  - 结果提取和因果关系识别
  - RDF 知识图谱构建
- 阶段 2
  - RDF 到 Coq 归纳类型转换
  - 高阶时序逻辑实现
  - 因果推断框架整合
  - 用于形式验证的证明助手兼容性

在第一阶段，我们使用三个增强策略在三个大型语言模型架构上，利用修昔底德的《伯罗奔尼撒战争史》（前 9 章）的历史文本。我们的知识来源包括包含 768 类和 3000 多个属性的 DBpedia 本体，以及用于古希腊地理位置的专门 LACRIMALit 本体 [27] 。外部知识检索包括 Wikidata 和 DBpedia SPARQL 端点，以及 ConceptNet API。我们使用的三个模型是 GPT-4o（大型商业模型）、Claude-3.5 Sonnet（替代大型架构）和 Llama 3.2（通过本地 Ollama 部署的 3.2B 参数模型）。所有模型都使用确定性设置（温度 =0）以保持一致性，并使用相同的分块参数。我们在这些模型上使用了三种变体，即基础生成、带知识增强的生成以及带知识增强和 RAG 的生成。

真实基础生成（控制）：纯粹的大型语言模型推理，不进行外部知识增强。为基线测量，所有 API、知识图谱、位置丰富和 RAG 功能被明确禁用。然而，基线使用了图 1 中所示的结构化提示。

知识增强（无 RAG）：在这里，我们使用知识整合但不使用 RAG。我们尝试实体提取，知识图谱检索查询到 Wikidata、DBpedia 和 ConceptNet。最后，我们尝试按照 LACRIMALit 本体 WikidataDBpedia 的顺序进行位置丰富，如果成功获得坐标则停止。

RAG 增强：这增加了一个检索增强生成层，该层将向量相似度与知识图谱增强结合在一起。实现中使用了 FAISS 向量存储和 HuggingFace 的 paraphrase-multilingual-MiniLM-L12-v2 嵌入，以及不同的参数 k。RAG 流程如图 2 所示。

System Prompt for Historical Event Extraction

You are extracting historical events from text using ONLY the information provided in the text chunk.

TEXT CHUNK { chunk_num } TO ANALYZE:
{ chunk }

ENTITIES FOUND: { entities }
LOCATIONS FOUND: { locations }

TASK: Extract historical events using text information and making REASONABLE INFERENCES.

REQUIREMENTS:

1. 仅提取文本块中明确定义的事件
2. 使用直接陈述在文本中的信息
3. 根据上下文线索做出合理推断
4. 从文本上下文推断坐标、国家、地区

OUTPUT FORMAT:

```
ste:Event { chunk_num } _1 a ste:Event ;
ste:hasType "event description" ;
ste:hasAgent "person/group" ;
ste:hasTime "time period" ;
ste:hasLocation "location name" ;
ste:hasLatitude "37.9838" ;
ste:hasLongitude "23.7275" ;
ste:hasCountry "Greece" ;
ste:hasRegion "Attica" ;
ste:hasLocationSource "inferred" ;
ste:hasResult "outcome" .
```

Fig. 1. 历史事件提取的 LLM 系统提示

3.1 实现与评估

我们的系统采用级联查询（本体 Wikidata DBpedia），并使用基于 MD5 的缓存和速率限制来减少 API 调用。评估结合了定量指标（事件覆盖率、坐标精度、RDF 完整性）和定性评估（语义深度和本体分类）。所有输出遵循标准化的 RDF/Turtle 格式，使用一致的属性: `ste:hasType`、`ste:hasAgent`、`ste:hasTime`、`ste:hasLocation`、`ste:hasLatitude/Longitude`、和 `ste:hasResult`。`ste:hasType`、`ste:hasAgent`、`ste:hasTime`、`ste:hasLocation`、`ste:hasLatitude/Longitude`、和 `ste:hasResult`，以实现增强策略的直接比较。

3.2 增强策略比较

我们的系统研究考察了不同的知识整合方法如何影响模型架构的抽取性能。我们使用相同的提示框架来评估三种增强策略，以确保公平比较。结果如表 1 所示。我们的分析揭示了显著的性能权衡，

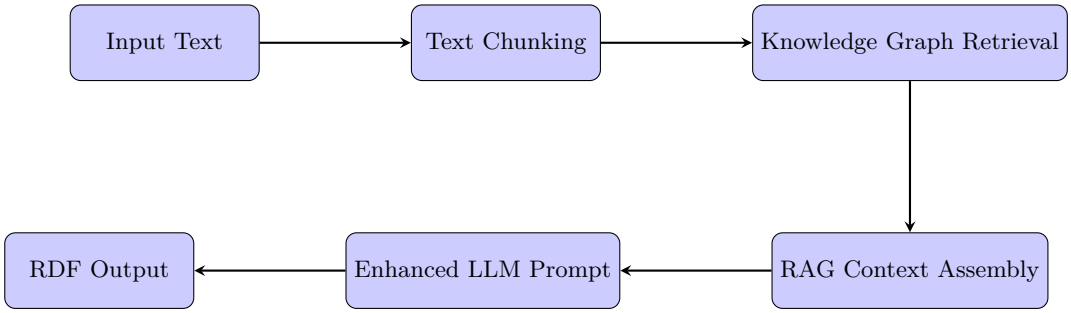

Fig. 2. RAG 管道

挑战了关于外部知识收益的传统假设。结果表明，增强策略在较强模型下有害，而在较弱模型下有帮助。基础生成始终以丰富的叙述细节和全面的历史范围实现了最大的事件覆盖（GPT-4: 36 个事件，Claude: 39 个事件），在语法上几乎生成了干净的 RDF（唯一缺少的是 xsd 格式，这在基础提示中缺失，但在其他架构中包含）。知识库和 RAG 一致地产生更少的事件，同时在文本中引入幻觉。例如，西伯塔战役的坐标在基础的 GPT-4 中正确归属于西伯塔，但在知识增强的 GPT-4 中错误地给出。一点研究显示，所使用的坐标属于安布拉基亚，西伯塔战役的参与城市之一。Claude-RAG 混淆了三个不同的时间层：古代事件（公元前 431 年战役、围城）、作者身份（修昔底德撰写的战争）、现代出版（Crawley 翻译、古腾堡计划更新），将所有这些都作为当代伯罗奔尼撒战争事件提取。此外，还包括元数据如 hasLLM: Claude4 作为 RDF 属性。

LLAMA 表现出最显著的性能变化，在复杂增强条件下表现出脆弱的失败模式。基础 LLAMA 仅产生 10 个格式较差的事件，建立了一个较低基线。简单的 RAG 配置（RAG 1-2）带来了显著的改进，将输出提升到 37-38 个结构更好的事件，表明知识增强可以有效弥补模型能力的不足。然而，在更高的复杂性水平上，这种改进表现出灾难性的失败：更复杂的 RAG 设置导致完全崩溃，RAG 3 仅产生 6 个事件，而 RAG 4 则没有实现任何正确的事件结构。这表明较弱的模型从受控的知识注入中受益，但缺乏处理复杂多源推理而不系统性崩溃的结构稳健性。

这些发现揭示了一个细致的反校准原则，该原则在两个关键维度上运作：增强强度和模型能力。最近的实证研究表明，更强的检索器反而导致更多的性能下降，性能展示出一种“倒 U 形模式”，即初步改进后出现急剧下降 [17] 。我们的结果将这一现象从检索扩展到更广泛的知识整合策略，揭示了模型强度与增强效益之间的关系遵循一种复杂的反校准，即更强的模型在最小增强下表现最佳，而较弱的模型则需要仔细限制的增强。

LLAMA 轨迹，即从基线失败（10 个事件）通过最佳性能（37-38 个事件）到灾难性崩溃（0 个事件）的过程，展示了知识整合中的脆弱性阈值。这与发现的结果一致，即较弱的模型更容易受到“困难否定”的影响，并且只能在特定复杂性限制内进行有效知识增强，否则将经历系统性崩溃 [17] 。从基线到简单 RAG 的改进表明，较弱的架构可以从受控知识注入中显著受益，有效弥补知识不足。然而，在更高 RAG 复杂性水平上的灾难性失败显示，这些相同模型缺乏在不发生语义崩溃的情况下处理多源推理的结构稳健性。

相反，像 GPT-4o 和 Claude 这样的较强模型表现出相反的模式：其复杂的参数知识和推理能力受到外部知识来源的干扰而不是增强。这支持了研究显示较强的模型在内部知识和检索上下文间表现出不同的关系，通常优先考虑外部信息即便它与其更准确的训练数据 [32] 相矛盾。在 GPT-4o 中的坐标错误归因和 Claude-RAG 中的时间混淆例证了知识增强如何引入噪声，从而压倒这些模型的模式识别能力，使它们倾向于从可靠的内部表述转向可能被损坏的外部来源。这种现象反映了理论框架，即知识整合增加认知负荷并引入语义漂移，特别对那些具有复杂内部知识结构的模型来说是一个问题，因为这些模型可能会被冲突的外部信息 [12] 所动摇。

Model	Configuration	Events	Precision	RDF	Issues
GPT	Base Generation	36	Excellent	Excellent	Missing xsd (due to prompt)
GPT	No RAG + KG	30	Good (some Hallucinations)	Excellent	Some empty coordinates
GPT	RAG + Multi KG 1	23	Good (some Hallucinations)	Excellent	Some empty coordinates
GPT	RAG + Multi KG 2	23	Good (some Hallucinations)	Excellent	some empty coordinates
GPT	RAG + Multi KG 3	20	Good (some Hallucinations)	Perfect	No syntax issues
GPT	RAG + Multi KG 4	26	Good (some Hallucinations)	Perfect	No syntax issues
Claude 4	Base Generation	39	Excellent	Excellent	Missing xsd (due to prompt)
Claude 4	No RAG + KG	20	Fair (hallucinations)	Good	Repeats event ids
Claude 4	RAG + Multi KG 1	10	Poor (Many Hallucinations)	Good	Missing coordinate fields
Claude 4	RAG + Multi KG 2	16	Poor (Many Hallucinations)	Good	Missing coordinate fields
Claude 4	RAG + Multi KG 3	19	Poor (Many Hallucinations)	Good	Missing coordinate fields
Claude 4	RAG + Multi KG 4	19	Poor (Many Hallucinations)	Good	Missing coordinate fields
LLAMA	Base Generation	10	Fair	Poor	Many formatting errors
LLAMA	No RAG + KG	19	Good	Fair	Some formatting errors
LLAMA	RAG + Multi KG 1	38	Good	Good	Some missing coordinates and repeated event ids
LLAMA	RAG + Multi KG 2	37	Good	Good	Some missing coordinates and repeated event ids
LLAMA	RAG + Multi KG 3	6	Poor	Poor	Many syntax errors
LLAMA	RAG + Multi KG 4	0	Bad	Bad	No correct event structure

Table 1. 模型性能比较

除了性能指标之外，RAG 在某些情况下似乎还会破坏已提取事件的语义准确性，并在与标准知识图评估时扰乱本体一致性。虽然一些提取的事件与 DBpedia 的子类型对齐（例如，SocietalEvent），但大多数特定领域的事件描述不在 DBpedia 的分类中（例如，标准本体中没有 WarEvent 或 MigrationEvent）。提示中提到的 dbp:SpecificEventType 显示了这一本体学上的缺口，即模型在遇到需要更细致语义区分的具有历史意义的事件时，通常会默认采用这一通用分类。然而，这些明显的不一致反映了一种根本的粒度不匹配，而不是真正的语义错误。RAG 持续发现特定领域的事件类型和关系，超出了通用本体的覆盖范围。虽然这些提取违反了 DBpedia 的限制，但它们通常捕捉了重要的、语义上连贯的历史区分，对于领域分析是至关重要的。当我们讨论将翻译程序转换为 Coq 时，这种子事件发现将会很有用。

最后，RAG 出错的其他情况是在需要考虑文本中更广泛的历史背景，但呈现的信息被混淆且不准确的情况，如你在下面的例子中所见：

表中 2 的示例不仅说明了时间合并问题，在该问题中修昔底德（记录战争的历史学家）被错误地识别为引发伯罗奔尼撒战争的代理者，而且还错误地将这场战争识别为涉及希腊人和大部分野蛮世界的历史上最大运动。这清楚地表明了 RAG 增强如何从根本上破坏历史人物和事件之间的语义关系，而不是使它们更加精确和丰富。

4 超越 RDF：实现高阶推理

虽然我们的事件抽取系统成功地从历史文本中生成了结构化的 RDF 知识图谱，但在面对复杂的分析需求时，这些表示形式遇到了基本的计算障碍。历史话语中固有的复杂时间依赖性、多层次因果网络和新兴叙事模式需要超越传统语义网框架计算边界的推理架构。

4.1 RDF/OWL 系统的计算限制

当代的 RDF/OWL 推理引擎在限于逻辑片段的情况下运行，这些片段对历史分析来说系统性地不足。这些系统被限制在可判定的第一阶逻辑子集内，从根本上限制了它们表达和验证复杂历史关系的能力。考虑从普拉塔亚围城战（公元前 429-427 年）到伯里克利政治衰退之间的级联因果

Table 2. RDF 事件提取示例：Claude-RAG 中的时间合并

Property	Value
`rdf:type`	`ste:Event, dbp:SpecificEventType`
`ste:hasType`	The outbreak of the Peloponnesian War
`ste:hasAgent`	Thucydides
`ste:hasTime`	431 BC
`ste:hasLocation`	Greece
`ste:hasLatitude`	39.0742 (`xsd:double`)
`ste:hasLongitude`	21.8243 (`xsd:double`)
`ste:hasCountry`	Greece
`ste:hasRegion`	Hellas
`ste:hasLocationSource`	wikidata
`ste:hasResult`	The greatest movement yet known in history, involving the Hellenes and a large part of the barbarian world
`ste:hasRAGContext`	yes

序列：持续的围城使雅典人口集中在城墙内，创造了加速瘟疫传播的条件，随后改变了政治情绪，削弱了伯里克利多年来的领导地位。虽然 RDF 可以将这些表示为具有简单因果谓词的离散事件（`:causedBy`），但无法形式化表达多步骤时间依赖关系或验证这种级联关系的逻辑一致性。

传统的 SPARQL 查询缺乏归纳推理模式的表达能力，这是历史分析所必需的。问题如“成功的军事行动是否与随后的政治稳定相关？”需要跨越时间序列的模式匹配和事件分布上的统计推断，这些能力超出了 SPARQL 的图模式匹配范式。此外，反事实推理要求模态逻辑结构，而这些在 RDF/OWL 规范中完全缺失。

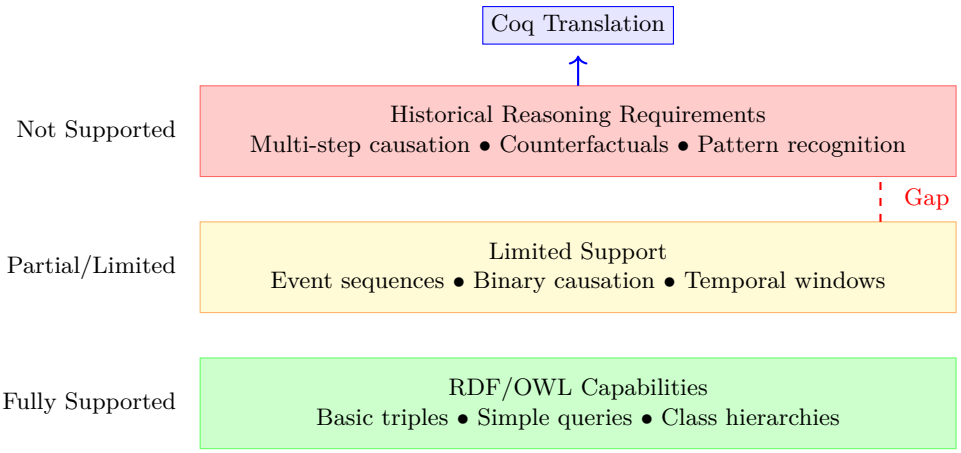

Fig. 3. RDF/OWL 能力与历史推理需求之间的差距。有限的 RDF 表达能力与复杂的历史分析之间的推理差距需要转化为形式验证系统。

4.2 形式翻译架构

我们开发了一个自动化翻译流程，将提取的 RDF/Turtle 表示转化为 Coq 证助工具的形式规范。该架构能够在进行高阶推理的同时，保留在 RAG 增强提取过程中捕获的语义丰富性，从而验证我们之前关于本体发现与幻觉的研究结果。

翻译系统由四个集成组件组成：(1) 一个 RDF 解析模块，从 Turtle 序列化数据中提取事件结构、时间关系和参与者角色；(2) 一个类型推理引擎，将 RAG 发现的事件类别映射到形式化的 Coq 归纳类型，表明 `MigrationEvent` 和 `WarEvent` 之类的细粒度分类代表合法的语义结构，而不是本体论上的违反；(3) 一个时态逻辑翻译器，使用依赖类型将时间关系转换为数学约束；以及 (4) 一个定理生成系统，产生关于历史因果关系的可验证命题，使复杂历史假设的形式化验证成为可能。

4.3 高阶推理能力

Coq 形式化在 RDF 框架中实现了不可能的分析能力。多步因果推理可以通过函数组合和依赖类型来表达：

```
Definition cascading_causation (e1 e2 e3 : Event) : Prop :=
  temporal_precedes e1 e2 / \  may_have_caused e1 e2 /\
  temporal_precedes e2 e3 / \  may_have_caused e2 e3 /\
  same_historical_context e1 e3.
```

这使得针对需要跨多个时间阶段的传递性推理的历史命题进行形式验证成为可能。归纳定义允许跨事件序列进行模式识别，而依赖类型则捕捉确保逻辑一致性的上下文约束。该系统可以正式地证明或反驳复杂的历史论点，将历史分析从解释性论述转变为严谨的逻辑研究。

至关重要的是，这种转换验证了我们的本体发现论点：违反 DBpedia 约束的 RAG 生成事件类型可以无缝集成到 Coq 的丰富类型系统中，从而展示其语义合法性。这些明显的本体不一致性代表了宝贵的领域特定知识结构，形式化推理系统可以利用这些知识结构来进行复杂的历史分析。

4.4 翻译过程

从 RDF/Turtle 表示到 Coq 规格的自动化翻译通过一个四阶段的管道操作，该管道系统地将提取的事件数据转变为形式逻辑构造，同时验证本体论的发现。

阶段 1：本体发现验证。翻译过程从分析从 RDF 三元组中提取的类层次结构开始，区分标准 DBpedia 类和 RAG 发现的事件类型。我们的验证算法使用语义指标（如“战争”、“迁移”、“围攻”的领域特定术语的存在）和上下文一致性（时间一致性、特定参与者、地理定位）来评估每个非标准类。这个验证步骤展示了明显的本体违规通常代表了合法的领域知识，而不是幻觉——例如，`ET_MigrationEvent` 代表了一种历史上准确的细粒度事件分类，超过了 DBpedia 的粒度。

阶段 2: 类型系统构建。翻译器通过将 RDF 类映射到带有正式子类关系的归纳类型来构建 Coq 的丰富类型系统。RAG 发现的事件类型无缝集成到这一层次结构中，验证了我们的论点，即这些代表连贯的语义结构。该系统为时间推理生成依赖类型，使得通过 Coq 的 Z 库在 BC 日期上进行正式算术运算成为可能：

```
Definition TimePoint := (Z * nat * nat)
Definition is_bc_event (e : Event) : bool :=
  match time e with
  | Some (y, _, _) => Z.ltb y 0
  | None => false
  end.
```

阶段 3：关系形式化。超出 RDF 表达能力的复杂事件关系转变为可正式验证的命题。通过函数组合和依赖类型进行多步骤的因果推理，使得在 SPARQL 的模式匹配范式内无法实现的级联历史因果关系的验证成为可能：

```
Definition cascading_causation (e1 e2 e3 : Event) : Prop :=
  event_before e1 e2 = true / \  may_have_caused e1 e2 = true /\
  event_before e2 e3 = true / \  may_have_caused e2 e3 = true /\
  same_historical_context e1 e3 = true.
```

第四阶段：定理生成。最后阶段生成形式化定理，以验证 RAG 结果的语义一致性，并支持复杂的历史推理。`rag_discoveries_coherent` 定理正式证明了具有正确时间、参与者和分类结构的 RAG 提取事件代表了有效的领域知识，而另一条定理则展示了领域特定分类如何增强因果推理能力。

这种翻译架构验证了我们的核心论点：在 RDF 评估中被识别为本体论违反的 RAG 发现事件类型，可能代表了有价值的语义创新，这些创新可以被形式推理系统利用，以进行超越传统语义网框架能力的复杂历史分析。

4.5 迈向实际应用：未来发展的基础

Coq 翻译为那些在传统 RDF/OWL 系统下不可能实现的应用奠定了基础，尽管在实际部署之前仍需进行重大开发。我们的系统展示了正式表示复杂历史关系的可行性，并在提取的事件结构中验证基本的逻辑一致性。虽然当前的应用需要 Coq 专业知识，但该架构证明了 RAG 发现的本体可以支持有关历史因果关系的形式推理。

主要贡献在于展示表面上的 RAG “幻觉”实际上是可以进行形式验证的合法本体论发现在。这样将评价标准从对现有知识图谱的符合性转变为识别特定领域的语义创新。我们的方法不是将细粒度的事件分类视为错误，而是通过形式数学证明验证其语义一致性。

该翻译架构开辟了几个有前景的研究方向。未来的工作可以开发领域特定的逻辑框架，这些框架可以捕捉历史推理模式，而不需要完全掌握 Coq。自动一致性检查工具对于数字人文学者来说是另一种有前景的方向，混合系统将统计相关性与形式逻辑约束相结合进行因果分析也是如此。此外，用户友好的界面最终可能让没有形式逻辑培训的历史学家能够使用形式验证功能。

有几个限制限制了其立即的实际部署。该方法需要大量的形式逻辑专业知识，这限制了在职历史学家的可访问性。该系统目前只能处理相对简单的历史关系，将其扩展到具有多个相互矛盾来源的复杂历史叙述仍然具有挑战性。翻译过程还需要手动验证所生成的 Coq 定义以确保其正确性。

这种将自动提取与形式验证相结合的方法，代表了朝着严格的计算人文学工具迈出了一步。虽然对于实际工作的历史学家来说尚不能立即部署，但该方法展示了现代自然语言处理技术可以以形式逻辑系统为基础，暗示未来工具开发的方向具有希望，这些工具最终可以为历史分析提供可扩展性和数学严谨性。

5 结论

我们对 RAG 配置的系统比较揭示了一个基本的见解：最佳 RAG 设计通过展示基础生成在没有外部增强的情况下通常优于中等和最大化检索策略，挑战了传统假设。发现纯推理生成在整体表现上优于增强的 RAG 配置，尤其是对于文献丰富的领域，这对该领域有重要意义。

这些发现挑战了普遍认定的更全面的检索必然会带来更好性能的假设。我们的分析表明，增强策略优化了不同的性能维度，而不是提供通用的改进。像 GPT-4o 和 Claude-3.5 这样的强大模型

在纯基础生成方面表现出色，能够生成具有丰富语义细节的全面历史覆盖，而外部增强则提供了互补但常常有害的效果：技术精度的提高以降低事件覆盖和引入幻觉为代价。

我们识别出的体系结构增强差异表明，模型容量从根本上决定了基线性能和增强敏感性。更大规模、经过广泛训练的模型拥有稳健的内部知识表示，从外部增强中选择性受益，同时可能牺牲叙述的广度。像 Llama 3.2 这样的较小模型需要外部知识支撑，但当增强复杂性超过其架构的稳健性时会表现出灾难性故障模式，表现出从具竞争力的性能到完全语义崩溃的高度差异性。

除了性能指标之外，我们的研究揭示了 RAG 系统不仅仅执行信息检索，而是进行本体发现。当根据像 DBpedia 这样的通用本体进行评估时，出现的细粒度事件类型虽然被视为偏离标准，但实际上可能代表合理的领域特定语义结构。事件例如 ET_Revolt 或 ET_MigrationEvent 超出了 DBpedia 的粒度，同时保持了历史准确性和语义连贯性。这将评估范式从遵循现有知识图谱重新定位为承认领域特定创新。

我们将其翻译成 Coq 证明，这些由 RAG 发现的事件知识库能够支持传统语义网框架中无法实现的高级逻辑推理。能够表达和验证多步骤的因果关系、使用公元前日期进行复杂的时间算术运算，并对历史因果关系进行形式证明，代表了超越 RDF/OWL 能力的重大进步。将 RAG 发现的事件类型成功集成到 Coq 的丰富类型系统中，通过数学证明验证了其语义的合法性。

该方法为计算人文学科的应用奠定了基础，这些应用将现代自然语言处理技术的可扩展性与形式逻辑系统的严谨性相结合。虽然当前的实施需要相当的形式逻辑专业知识，但该架构证明了自动本体构建可以以数学验证为基础，开辟了未来工具开发中有前景的方向。

这些发现表明，RAG 系统设计应优先谨慎评估是否有必要进行外部增强，配置选择应依据领域特定的特征和模型能力。而对于经典的历史文本，LLM 的内置知识证明是相当全面的，这表明外部增强的价值主张需要进行谨慎的领域特定评估，而不是盲目应用。

未来的工作应探索跨领域和历史时期的泛化，研究结合基础生成的全面覆盖和目标增强的精确性的混合方法，并开发可访问的形式验证接口。将本体发现验证与自动推理相结合，为计算语言学和数字人文学科提供了有前途的方向。

6

致谢 作者对来自亚马逊的资助表示感谢（项目：用于增强自然语言处理的神经-符号集成 (NIELS)）。

References

1. Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Retrieval-based language models and applications. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts , pages 1–50, 2023.
2. Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. In Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023) , pages 70–98, Toronto, ON, Canada, 2023. Association for Computational Linguistics.
3. R. C. Barron, V. Grantcharov, S. Wanna, M. E. Eren, M. Bhattarai, N. Solovyev, G. Tompkins, C. Nicholas, K. Ø. Rasmussen, C. Matuszek, et al. Domain-Specific Retrieval-Augmented Generation Using Vector Stores, Knowledge Graphs, and Tensor Factorization. In 2024 International Conference on Machine Learning and Applications (ICMLA) , pages 1669–1676, 2024.
4. Bernardy, J.-P., Chatzikyriakidis, S.: Applied temporal analysis: A complete run of the FraCaS test suite. In: Proceedings of the 14th International Conference on Computational Semantics (IWCS). pp. 11–20 (2021)

5. Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning , pages 2206–2240, 2022.
6. Jean-Philippe Bernardy and Stergios Chatzikyriakidis. Applied temporal analysis: A complete run of the FraCaS test suite. In Proceedings of the 14th International Conference on Computational Semantics (IWCS) , pages 11–20, 2021.
7. Kartik Sharma, Peeyush Kumar, and Yunqing Li. Og-rag: Ontology-grounded retrieval-augmented generation for large language models. arXiv preprint arXiv:2412.15235 , 2024.
8. Stergios Chatzikyriakidis and Zhaohui Luo. Proof assistants for natural language semantics. In Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996–2016) 9th International Conference, LACL 2016, Nancy, France, December 5-7, 2016, Proceedings 9 , pages 85–98. Springer, 2016.
9. Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval , pages 719–729, 2024.
10. Xinya Du and Claire Cardie. Event extraction by answering (almost) natural questions. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 671–683, 2020.
11. Xinya Du and Heng Ji. Retrieval-augmented generative question answering for event argument extraction. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , pages 4649–4666, 2022.
12. Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. Nature , 630(8017):625–630, 2024.
13. Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 , 2023.
14. Shiming He, Yu Hong, Shuai Yang, Jianmin Yao, and Guodong Zhou. Demonstration retrieval-augmented generative event argument extraction. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) , pages 4617–4625, 2024.
15. Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. ACM Computing Surveys , 54(4):1–37, 2021.
16. Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics , pages 874–880, 2021.
17. Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. arXiv preprint arXiv:2410 , 2024.
18. Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 6769–6781, 2020.
19. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al. DBpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web , 6(2):167–195, 2015.

20. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Advances in Neural Information Processing Systems , volume 33, pages 9459–9474, 2020.
21. Hao Li, Yanan Cao, Yubing Ren, Fang Fang, Lanxue Zhang, Yingjie Li, and Shi Wang. Intra-Event and Inter-Event Dependency-Aware Graph Network for Event Argument Extraction. In Findings of the Association for Computational Linguistics: EMNLP 2023 , pages 6362–6372, Singapore, 2023. Association for Computational Linguistics.
22. Jasper Linders and Jakub M. Tomczak. Knowledge graph-extended retrieval augmented generation for question answering. arXiv preprint arXiv:2504.08893 , 2025.
23. Jiawei Liu, Chunyu Kit, and Gina-Anne Levow. RAG-based temporal reasoning for event relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence , pages 13245–13253, 2023.
24. Haoran Luo, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and others. HyperGraphRAG: Retrieval-Augmented Generation with Hypergraph-Structured Knowledge Representation. arXiv preprint arXiv:2503 , 2025.
25. Yubo Ma, Zehao Wang, Yixin Cao, and Aixin Sun. Few-shot event detection: An empirical study and a unified view. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics , pages 2845–2857, 2023.
26. Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms. arXiv preprint arXiv:2312.05934 , 2023.
27. Maria Papadopoulou, Christophe Roche, and Eleni-Melina Tamiolaki. The LACRIMALit Ontology of Crisis: An Event-Centric Model for Digital History. Information , 13(8):398, 2022.
28. Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web , 8(3):489–508, 2017.
29. W. Su, Y. Tang, Q. Ai, Z. Wu, and Y. Liu. DRAGIN: Dynamic Retrieval Augmented Generation based on the Information Needs of Large Language Models. arXiv preprint arXiv:2403.10081 , 2024.
30. Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. RA-NER: Retrieval augmented NER for knowledge intensive named entity recognition. Amazon Science Publications , 2024.
31. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Retrieval-augmented event argument extraction. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics , pages 2156–2168, 2023.
32. Kevin Wu, Eric Wu, and James Zou. How faithful are RAG models? Quantifying the tug-of-war between RAG and LLMs' internal prior. arXiv preprint arXiv:2404 , 2024.
33. Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In Advances in Neural Information Processing Systems , volume 35, pages 14951–14964, 2022.
34. Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Retrieval-augmented generation across heterogeneous knowledge. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pages 52–58, 2022.
35. In the real Dai, Chen Luo, Zhen Li, Xianfeng Tang, Hanqing Lu, Rahul Goutam, and Haiyang Zhang. RA-NER: Retrieval augmented NER for knowledge intensive named entity recognition. In The Second Tiny Papers Track at ICLR 2024 , 2024.
36. Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv:2402.19473 , 2024.
37. Bingfeng Zhou, Yangqiu Song, and Kam-Fai Wong. Event temporal relation extraction based on retrieval-augmented on llms. arXiv preprint arXiv:2403.15273 , 2024.