LexChronos：印度法学中结构化事件时间线抽取的智能体框架

Anka Chandrahas Tummepalli，Preethu Rose Anish

TCS Research ankachandrahas.t@tcs.com,
preethu.rose@tcs.com

arXiv:2603.01651v1 [cs.CL] 2 Mar 2026

摘要

理解和预测司法结果，需要对法律文档进行细致入微的分析。传统方法将判决和诉讼程序视为非结构化文本，限制了大语言模型（大语言模型）在摘要生成、论点生成和判决预测等任务中的效用。我们提出*LexChronos*，一种智能体框架，通过迭代方式从印度最高法院的判决中抽取结构化事件时间线。*LexChronos* 采用双智能体架构：LoRA 指令微调抽取智能体负责识别候选事件，预训练的反馈智能体则通过置信度驱动的循环对其进行评分和优化。为解决印度法律事件数据集匮乏的问题，我们运用逆向工程技术，借助 DeepSeek-R1 和 GPT-4 构建了包含 2000 个样本的合成语料库，并生成黄金标准事件标注。针对该合成真实标注，我们的流程取得了 0.8751 的基于 BERT 的 F1 分数。在法律文本摘要的下游评估中，GPT-4 在 75% 的案例中更倾向于结构化时间线，而非非结构化基线，表明结构化时间线能够提升对印度法学的理解和推理能力。本研究通过利用法律事件的结构化表示，为印度背景下未来的法律人工智能应用奠定了基础，例如判例映射、论证综合和判决预测建模。

关键词

智能体人工智能、事件抽取、法律文档分析、印度法学。

1 引言

高风险情境下的法律推理不仅要求精准解释，还要求对复杂案件叙事形成结构化理解。这一要求凸显了事件抽取的关键作用——识别并结构化关键事实、参与者及时间关系的过程（Xian 等人，2024年）。尽管至关重要，由于司法文书固有的复杂性和非结构化特征，事件抽取仍是法律人工智能系统面临的一项持续挑战，尤其是在文档层面，必须跨越冗长文本跟踪事件。

在全球范围内，人工智能已从理论应用发展为法律行业的实用工具，可支持各类任务

版权所有 © 2026年，人工智能促进协会（www.aaai.org）。保留所有权利。

例如转录、翻译、案件管理（印度政府新闻信息局，2025年）、智能排期以及用于减少积案的预测分析。大语言模型正越来越多地应用于法律研究、文书起草和分析（Deeks 和 Hollis，2025年）。然而，从法院判决中提取结构化事件时间线仍鲜有探索——尤其是在印度司法机构这一全球规模最大的司法体系之一的复杂环境中。截至2025年，仅印度最高法院就有86,742 起未决案件（Rohit Singh，2025年）。

印度法律文本面临着独特挑战：成文法语言密集、涉及普通法、包含宪法引文，并存在程序依赖关系。传统自然语言处理技术往往无法捕捉这些文档中蕴含的时间、因果和层级结构。尽管句子级事件抽取已取得进展，但文档级抽取仍是瓶颈，因为它需要进行共指消解、跨段落实体跟踪以及时间因果关联（Huang 和 Peng，2021年）。缺乏结构化表示严重限制了判例映射、论点生成和判决预测等下游任务。

尽管近年来面向法律应用的大语言模型取得了进展（Tiwari 等，2024年），提升了文本理解能力，但大多数方法仍将法律文档视为整体文本块，忽视了支撑法律推理的细致事件结构。尽管印度最高法院已采用人工智能来提高行政效率，从判决中全面抽取事件仍是一个尚未充分探索、具有变革潜力的领域。

本文针对两个关键缺口：

（1）大语言模型在从非结构化法院文书中抽取和构建事件方面存在局限。（2）缺少面向印度最高法院判决、公开可用且按事件级别标注的数据集，妨碍了可复现性和基准测试。

为弥补这些缺口，我们提出了*LexChronos*，一个用于从法律文本中迭代抽取结构化事件时间线的智能体框架。我们的双智能体流水线由以下部分组成：

1. 抽取智能体（LoRA指令微调大语言模型 $<4B$ 参数）提出候选事件，包括时间戳、描述、法官和先例引用。

2. 反馈智能体（大语言模型 $< 4B$ 参数）评估候选项——

分配置信度分数，并通过由停止标准驱动的循环优化输出。这种元认知反馈确保语义连贯性和事实准确性，使时间线逐步收敛至高置信度状态。

为建立基准，我们使用 DeepSeek-R1（Guo等（2025））和 GPT-4（Achiam等（2023））构建了一个包含 2000 份最高法院判决的合成语料库，并为其提供了黄金标准事件标注。

我们的主要贡献如下：

- 一种双智能体迭代优化框架，用于从非结构化法律文本中提取结构化事件时间线。
- 一个专为事件级法律自然语言处理任务打造的合成标注数据集，其中包含 2000 份印度最高法院判决。

通过引入结构化事件时间线作为中间表示，*LexChronos* 提升了法律推理的透明度和可靠性，为先例映射、论点生成和判决预测等高级工具的发展铺平了道路。这项工作为在印度司法系统中实现智能且可靠的人工智能集成奠定了基础。

本文其余部分安排如下：第 2 节回顾相关工作，第 3 节介绍数据集构建，第 4 节阐述方法，第 5 节介绍实验与评估结果，第 6 节提供数据和代码可用性信息，第 7 节讨论有效性威胁，第 8 节以结论和未来工作作结。

2 相关工作

事件抽取在将非结构化文本转化为结构化表示方面发挥着关键作用，为摘要生成、时间线生成和决策支持等下游任务提供支持。在不同领域中，从生物医学文献到金融披露、新闻媒体和法律文件，事件抽取不断发展，以应对嵌套结构、长距离依赖和语义歧义等领域特有的挑战。本节将从三个方面回顾关键进展：跨领域的事件抽取、法律领域的专门方法，以及针对印度法律体系的新兴研究。

2.1 跨领域的事件抽取

在生物医学领域，为处理嵌套事件结构和长距离依赖，BEESL模型（Alan等，2020）将生物医学事件抽取重新定义为统一的序列标注任务，采用基于BERT的编码、多标签解码和多任务学习，从而实现对结构复杂性的端到端处理。此外，结合依存路径嵌入和多种特征类型的卷积神经网络，已被证明在基于流水线的事件与关系抽取中效果显著（Bj¨orneand Salakoski，2018）。

在金融领域，公司公告等文档因篇幅较长、语义密度较高而带来挑战。DocFEE数据集（Chen 等，2025年）支持文档级中文金融事件抽取，而矩阵分块方法（MACK）（Huang 等，2024年）则引入了容错机制，以便从原始文本中稳健地抽取信息。同样，在新闻媒体领域，混合方法结合了基于模式的启发式方法、机器学习模型和词嵌入，已被用于在实时风险评估场景中保持较高的触发词准确率（Han、Hao 和 Huang，2018）。

2.2 法律领域的事件抽取

法律事件抽取已从基于规则的系统发展到由先进语言模型驱动的框架。早期研究（Sierra 等人，2018）运用词性标注和语法模式匹配，从墨西哥西班牙语法律文本中抽取核心事件属性。（Filtz 等人，2020）在欧洲人权法院数据集上对条件随机场、深度学习和微调后的 BERT 变体进行了基准测试，结果表明，条件随机场擅长行为主体识别，而 BERT 变体在时间和程序分类方面表现更优。（Xian 等人，2024年）推出了 DLEE，这是首个大规模中文法律文档级事件抽取数据集。他们采用半自动标注流程，结合专家设计的模式、触发词词典和基于 BERT 的问答模型，以确保标注细致且高质量。

（Xu，2024年）结合 BERT、双向长短期记忆网络和条件随机场架构，从经济类法律文本中抽取事件实体，取得了稳健的跨体裁表现。近期，大语言模型的进展使更灵活、更具上下文感知能力的抽取成为可能。（Yue 等人，2024年）提出了一种用于刑事裁判观点生成的协作框架，利用大语言模型构建事件叙事，为判决说理提供支持。

基于图的方法，例如（Zhao、Zhao 和 Mao2024年）提出的方法，使用事件实体抽取构建司法知识图谱，从而提升文档的可解释性和法律推理能力。

2.3 印度法律事件抽取

在印度法律语境中，(Naik, Patel, and Kannan 2023) 开发了一种基于 spaCy 的命名实体识别模型，用于从印度法律文书中提取法院名称、申请人和法令等关键实体。(Kalamkar et al. 2022) 推出了一个大型标注语料库，包含来自印度法院判决的 46,545 个法律命名实体，涵盖 14 种细粒度类型。他们基于该数据集开发并发布了一个基线命名实体识别模型，以支持关系抽取和知识图谱构建等下游法律自然语言处理任务。(Tiwari et al. 2024) 推出了 AALAP，这是一种基于专门法律语料微调的 Mistral 7B 模型，可用于法律争点生成、论点生成和事件时间线构建等任务。尽管 AALAP 在若干任务上的表现优于 GPT-3.5，但在摘要生成和宪法问答方面表现不佳。(Hussain and Thomas 2024) 对包括 Mistral 和 Gemma 在内的大语言模型进行了比较评估，考察它们从印度判例法中提取司法实体的能力。他们对这些模型进行了微调和测试，以评估其在自动执行与法律自然语言处理相关的结构化信息抽取任务方面的效果。Legal Sarathi (Shivananda et al. 2024) 等系统将大语言模型与机器学习算法相结合，从非结构化法律文档中提取关键事件及参与者。其检索增强生成（RAG）框架和 Streamlit 界面支持实时的基于查询的信息检索与决策。

判决组成部分	***LES*** 覆盖范围	覆盖范围的依据
事实	时间戳、事件	事实要素编码为**事件**，并附有相应的**时间戳**用于确立案件时间线。
争议事项	事件	争议事项的提出是一项关键的程序性**事件**，它决定了后续法律分析的重点。
申请人的论点	事件，先例	论点构成程序性**事件**；其依据很大程度上依赖**先例**或法条引用。
被申请人的论点	事件，先例	与申请人的主张类似，这些都是有时间标记的程序性**事件**以**先例**为依据。
法律分析	事件、先例	法律适用是一项司法**事件**，并且本质上与所讨论的**先例** 相关联。
判例分析	事件、先例、法官	这一分析性**事件**明确涉及所引用的**先例**以及进行分析的权威**法官**。
法院的推理	时间戳、事件、法官	推理过程构成了决定性的**事件**，由**法官**，一直到命令的最终**时间戳**。
结论	时间戳、事件、法官	最终命令是具有决定性的**事件**，由作者——权威**法官** 在最终**时间戳**作出。

表 1：判决核心组成部分与 LES 的映射

尽管取得了这些进展，针对印度司法语境的资源仍存在显著缺口。现有研究往往聚焦于范围狭窄的事件类别，缺乏全面覆盖。尤为重要的是，目前没有公开可用的数据集能够涵盖印度最高法院文书中的全部事件类型。

为解决这一问题，我们提出了*LexChronos*，这是一个智能体框架，旨在从印度最高法院判决文书中迭代抽取结构化事件时间线。我们使用 DeepSeek R1 和 GPT-4 通过逆向工程技术构建了包含 2,000 份文书的合成语料库，并生成高质量的黄金标准事件标注。*LexChronos* 采用双智能体架构：LoRA 指令微调抽取智能体负责识别候选事件，反馈智能体则利用置信度驱动停止准则对其进行优化。

3 数据集构建

为解决印度法律领域缺乏公开可用的事件级标注资源这一严峻问题，我们构建了一个合成数据集，其内容专为适应印度最高法院判决的结构和语义而设计。该数据集通过一条逆向工程流水线生成，旨在对法律程序生成高保真度的结构化表示。

数据集构建流水线包含三个关键阶段：案件类别选择、事件时间线生成和判决文本生成。

3.1 案件类别选择

印度最高法院的案件类别是依据案件所涉法律争议、争议事项、合议庭组成（合议庭组成）或其他相关考量因素，对案件进行的不同分类（Supreme Court of India 2025b）。数据集构建流程首先从印度最高法院官方案件分类系统（ Supreme Court of India 2025a）所定义的25个案件类别中，随机选取一个类别。

这些类别经过精心挑选，涵盖案件量大、社会影响显著且专业性强的领域，既包括刑事法和民法等基础领域，也包括网络法和知识产权等新兴领域。这种多样性确保合成数据集不会产生狭隘的算法偏差，并能反映印度法律案件在程序和主题方面的广泛性。案件类别的完整列表见附录A。

3.2 事件时间线生成

选定案件类别后，下一步是生成法律事件的结构化时间线。本节将介绍事件时间线生成的流程及其验证。

结构化事件时间线构建

对于每个选定的案件类别，我们使用提示词（$Prompt_E$），通过DeepSeek-R1和GPT-4生成事件的结构化时间线。

$$E = Prompt_E(case\ category) \tag{1}$$

每个输出 E 都是一个由事件对象构成的JSON数组，其中每个对象包含四个属性：

- 时间戳：事件发生的时间或日期。
- 事件：对事件的叙述性描述。
- 法官：主持审理的法官姓名（如适用）。
- 先例：事件中引用的法律先例。

这一四属性模式（以下称为 Lex-Chronos 事件模式——LES）是通过对印度最高法院判决文书进行归纳分析得出的。通过考察司法推理的结构组成部分——如时间演进、司法主体、事实和援引先例——我们将其抽象为 LES 中的四项核心属性。这使判决的法律史能够以细粒度、机器可读的形式呈现。

模型	判决平均长度	每个案件的平均事件数	每个案件的平均先例数	唯一词汇量
DeepSeek-R1	1010.96	27	6	34623
GPT-4	552.92	19	3	17080

表2：DeepSeek R1 与 GPT-4 判决的结构与语言比较

判决核心组成部分与判决文本
事实
“争议起源可追溯至2020年1月，当时 Proxima Inc. 收购了 Zenith Corp 的大量股份，由此引发了有关公司治理和少数股东权利的争议。”
争议事项
“……讨论的主要争议事项围绕公司治理、少数股东的权利，以及对公司决策进行司法监督的许可范围。”
申请人的论点
“2020年2月，Zenith Corp 的少数股东指控 Proxima Inc. 的董事会成员存在压迫与管理不善行为，并因此提起申诉。……”
被申请人的论点
“Proxima Inc. 辩称，其行为符合《2013年公司法》，并援引先例——Furlong Steel Ltd. v. Cherry Steel Corp. 5 SCC 739，以证明其董事会决策的合法性。”
法律分析
“经仔细审议所提交的论点和证据，我们认为，保护少数股东的必要性——并确保公司治理透明至关重要。”
判例分析
“Proxima Inc. 因禁令而不服，于2022年3月向高等法院提起上诉，并援引了以下先例：Bright Solar Ltd. v. Gloomy Electric Co. 9 SCC 600……”
法院的推理
“根据证据，尤其是法务审计报告所揭示的情况，国家公司法法庭的认定似乎依据充分，该报告揭示了严重的财务管理不善。”
结论
“鉴于上述情况，我们维持国家公司法法庭的判决，确认对 Proxima Inc. 董事会的重组以及对受影响少数股东所判给的赔偿。因此，上诉予以驳回……”

表3：合成判决文本的结构组成映射

LES 验证

为验证 LES 的充分性，我们将其与印度最高法院判决中通常包含的八个核心组成部分进行对照，这些组成部分由 IndianKanoon（Indian Kanoon 2025年）等法律信息检索系统所识别。表1展示了这一映射，表明四属性模式能够有效涵盖全面法律理解所需的全部实质性法律内容——事实、推理和结论。这证实了该模式对于事件级抽取的充分性及其不可约性。

3.3 判决文本生成

在结构化事件时间线就绪后，我们继续生成相应的判决文本。此步骤确保数据集同时包含结构化标注和真实的法律叙事。

判决文本构建

使用结构化事件时间线 E，第二个提示（$Prompt_J$）会生成一份合成判决文书：

$$J = Prompt_J(E) \tag{2}$$

$Prompt_J$ 其生成依据是从 IndianKanoon 等平台获取的真实判决中提炼出的模式。这确保生成的判决文本（J）能够仿真实际最高法院判决的语言复杂性、结构惯例和格式要求，这对下游法律推理任务至关重要。

通过将整个流程重复执行 2,000 次，我们构建了完整的合成数据集：

$$D = \{(E_i, J_i)\}_{i=1}^{2000} \tag{3}$$

该数据集将经过丰富标注的事件时间线与所有 25 种案件类别中相应的判决文本配对。样本规模设定为 2,000 条，旨在充分涵盖各类别在语言和结构上的差异，同时确保生成和下游处理在计算上可行。DeepSeek-R1 生成了 1,000 条样本，GPT-4 则生成了其余 1,000 条。选择 DeepSeek-R1 是因为它擅长生成结构化输出且逻辑一致；选择 GPT-4 则是因为它在长篇法律文本生成方面语言流畅、善于忠实呈现修辞（Guo等（2025）；Achiam等（2023））。表 2 比较了 DeepSeek-R1 和 GPT-4 生成的判决在关键结构和语言特征方面的差异。

验证为验证合成语料库的结构保真度，我们详细拆解了随机选取的一份判决文本 J，并展示其各组成部分。表 3 展示了合成文档中全部八个核心判决组成部分的完整性与清晰度。这证实了嵌入 $Prompt_J$ 中的生成约束能够成功复现真实的司法修辞和组织结构。

$Prompt_E$和 $Prompt_J$均采用零样本提示，并结合角色提示（Wang 等，2024）和风格提示（Lu 等，2023）加以增强。这些提示可在 GitHub 仓库中获取（Tummepalli 2025）。该策略确保结构化事件标注与生成的判决文本保持一致，发挥约束满足机制的作用，从而保证输出的高质量。通过整合多样的案件类别、严格定义的事件模式以及双模型生成策略，这一合成数据集可支持法律自然语言处理模型的稳健训练与评估，尤其适用于印度法律语境下的事件抽取任务。图1展示了用于构建数据集的完整流水线。

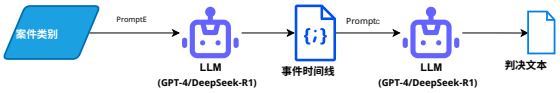

图1：金标准事件时间线数据集的逆向工程流水线

4 方法论

*LexChronos*采用双智能体架构，由抽取智能体和反馈智能体组成，二者在预设停止标准的约束下，通过迭代优化循环协同工作。这种协作设计通过渐进式验证与反馈，确保事件时间线抽取得到系统性改进。

4.1 事件时间线抽取

*LexChronos*遵循结构化的迭代流程，对每份判决文书 J 逐步进行精炼。一个通过 LoRA 在我们的合成数据集上训练的指令微调开源大语言模型（$<= 4B$ 个参数）充当抽取智能体。给定一份判决文书 J，抽取智能体会生成一组初始候选事件$E_0 = (e_1, e_2, ..., e_n)$，每个事件 e_i均包含时间戳、描述、法官和判例信息，形成一个初步的

一个预训练语言模型（$< 4B$ 参数）充当反馈智能体，根据 J 评估 E_0，并从多个质量维度生成反馈 F_0。该过程迭代如下：

$$E_0 = ExtractionAgent(J) \tag{4}$$

$$F_0 = FeedbackAgent(J, E_0) \tag{5}$$

$$E_{i+1} = ExtractionAgent(J, E_i, F_i) \tag{6}$$

$$F_{i+1} = FeedbackAgent(J, E_{i+1}) \tag{7}$$

其中，E_i、F_i 表示细化过程中的中间阶段。E_{i+1} 和 F_{i+1} 将是最后一对细化步骤，其中，如果满足停止条件，E_{i+1} 将是最终提取的事件时间线。

当两种停止条件（耐心上限和容差阈值上限）中的任意一种满足时，迭代过程即告终止。

•当连续 3 次迭代的置信度分数均未超过已达到的最高分数 S_best 时，将触发**耐心上限**:

$$\forall j \in \{i-2, i-1, i\} : S_j \leq S_{best} \tag{8}$$

• 当置信度分数连续 3 次迭代保持不变时，将触发**容差阈值上限**:

$$S_{i-2} = S_{i-1} = S_i \tag{9}$$

实证评估表明，迭代三次在性能与计算成本之间取得了最佳平衡。该流程会生成经过完善的事件时间线 E_k，并附上最终反馈 F_k，通过智能体协作确保质量得到系统性提升。

反馈智能体从七个维度对提取的事件进行评分（$0.00 - 1.00$ 量表）：叙事相关性 （与案件叙述的契合度）、时间准确性 （日期是否正确）、时间顺序流畅性 （事件排序是否合乎逻辑）、事件细节 （描述性细节是否充分）、重复项 （是否检测到重复内容）、角色识别 （参与者角色是否明确）以及置信度分数（整体提取质量）。评分还附有文字批评意见，指出需要改进之处。置信度分数是主要的收敛信号。通过这种系统化的阈值判定，可以确认输出质量已达到精确率和召回率的最高水平，使系统能够有把握地停止，而不是继续进行成本高昂、产出有限的生成。通过这种结构化的迭代机制，最终提取的时间线 E_k 在传递给下游应用之前，能够具备尽可能高且可验证的可靠性。在初始提取和完善阶段，我们采用了零样本提示以及角色提示和风格提示。在生成反馈时，我们采用了结合自我批评（Huang 等，2023）和线性量表（Liu 等，2023）技术的零样本提示。整体框架架构和迭代完善流程如图 2 所示。这些提示词已发布在 GitHub 仓库中（Tummepalli 2025）。

4.2 下游任务评估

为展示结构化事件时间线表示在法律人工智能应用中的实用价值，我们评估其对一项关键下游任务——判决摘要生成——的影响。我们从第 3 节介绍的合成数据集中随机选取 200 份判决样本，记为$J = (J_1, J_2, ... J_{200})$，并为每份样本配对相应的真实事件时间线 $E = (E_1, E_2, ... E_{200})$。

我们使用两种输入表示进行比较摘要实验。

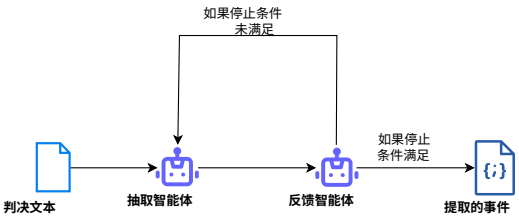

图 2：LexChronos 框架：用于迭代抽取与反馈驱动的优化

• 非结构化方法：将原始判决文书作为输入，要求大语言模型从冗长、杂乱的文本中提取关键论点和裁决结果。

$$S_{unstructured} = LLM(J_j) \tag{10}$$

• 结构化方法：生成的事件时间线作为输入，提供简洁且按时间顺序排列的叙述。

$$S_{structured} = LLM(E_j) \tag{11}$$

两种方法均使用相同的开源语言模型生成摘要，以确保公平性，并隔离结构化表示对摘要质量的影响。摘要提示采用零样本提示与角色与风格提示相结合的方式。我们采用零样本提示生成摘要，以确保不同模型之间的公平性和可复现性，并将输入表示（结构化与非结构化）的影响与模型适配的影响区分开来。虽然少样本或经过微调的模型可能会提升性能，但我们的目标是独立评估结构化表示的贡献，不受适配因素影响，因此零样本提示最适合用于受控评估。

评估时，我们采用 GPT-4 作为评审大语言模型，并使用成对比较协议：评估器会收到原始判决文本以及两份候选摘要，然后判断哪一份更符合司法质量标准。选择 GPT-4 是因为它在摘要评估方面已表现出与人类评估者高度一致；研究也指出，成对比较是最有效的评估方法（Liu 等，2024）。鉴于这种高度相关性，本研究未纳入人工评估，因为 GPT-4 是一种可靠且经济高效的替代方案，适用于大规模实验。

我们的评估方案依据一份详细的核对清单，该清单包含八项权重相同的标准——叙述简洁流畅（法律脉络清晰）、简洁性（精炼而完整）、结果清晰度（明确给出最终裁决）、聚焦关键要素（突出核心争议事项）、引用精简（仅保留必要引文，避免法律术语）、便于从业者阅读 （采用通俗、专业的语言）、细节平衡 （仅提供必要背景）、时间顺序合理（按逻辑顺序编排）。这些标准共同体现了法律摘要在准确性、可读性和实用性之间所需的平衡。法律摘要所需的实际效用。评审大语言模型会选出满足大多数标准、同时忠实于原判决文本 J 的摘要。为减少偏差，评估提示词采用零样本提示并结合自我批评，要求 GPT-4 以可理解的理由说明其偏好。所有摘要生成和评估提示词均公开于我们的 GitHub 仓库（Tummepalli 2025）。

5 实验与评估结果

我们分三个连续阶段评估*LexChronos* ：（i）对抽取智能体进行指令微调，（ii）评估反馈智能体对抽取质量的贡献，以及（iii）衡量结构化事件时间线对法律判决摘要生成的影响。所有实验均使用第3节所述的合成数据集。

5.1 抽取智能体性能

为确定最有效的抽取智能体，我们在合成语料库的训练集（$n_{train} = 1600$）上对八个开源语言模型（每个参数量均不足40亿）进行了指令微调，并保留 $n_{test} = 400$ 用于评估。所有模型均采用基于 LoRA 的指令微调进行微调，该方法仅更新一小部分参数，从而确保计算效率。候选模型包括：Llama 3.2 3B Instruct（Meta 2024年b）、Llama 3.2 1B Instruct（Meta2024年a）、Gemma2 2B IT（Google 2024年）、Gemma3 1B IT（Team等，2025年）、DeepSeek R1 Distill Qwen 1.5B（Guo等（2025））、Phi-4-Mini-Reasoning（Xu等，2025年）、Qwen 2.5 3B（Tea m 2024年）以及Qwen 3 4B（Yang等，2025年）。

微调后，我们使用基于 BERT 的精确率、召回率和 F1 分数（Zhang et al. 2019），将每个模型在留出的测试集上的表现与黄金标准事件标注 E_i 进行比较。选择基于 BERT 的指标在法律领域至关重要：ROUGE 或 BLEU 等传统指标依赖词汇重叠，往往无法捕捉法律语言中的语义细微差别；在法律语言中，细微的措辞选择蕴含着重要的语境和法律含义（Kaster、Zhao 和 Eger，2021）。BERTScore 利用上下文嵌入，衡量抽取出的事件与真实标注之间的语义连贯性和等价性，因此与人工判断的文本忠实度具有更强的相关性。对于复杂的长篇法律文档，语义等价性远比词元级匹配更能可靠地体现抽取质量，因此基于 BERT 的指标是更优的选择。

表4展示了结果，其中 Llama 3.2 3B Instruct 在所有候选模型中取得了最高的基于 BERT 的 F1 分数（0.8390）。这一优异表现表明，尽管参数数量较少，针对推理进行优化的架构设计仍能显著提升结构化信息抽取效果。因此，Llama 3.2 3B Instruct 被选为最终的抽取智能体。

5.2 反馈代理的贡献

基于表现最佳的抽取智能体，我们通过将多个反馈模型整合到迭代优化循环中，评估反馈智能体的作用，其中

LLM	***BERT*精确率**	***BERT*召回率**	***BERT F1*值**
Llama 3.2 3B Instruct	**0.8450**	**0.8336**	**0.8390**
Gemma 3 1B IT	0.8237	0.7992	0.8113
Gemma 2 2B IT	0.7809	0.7957	0.7880
Llama 3.2 1B Instruct	0.6058	0.6054	0.6052
DeepSeek R1 Distill Qwen 1.5B	0.5548	0.5135	0.5276
Qwen 2.5 3B	0.5628	0.4467	0.4752
Phi 4 Mini 推理版	0.4802	0.2442	0.3048
Qwen 3 4B	0.4472	0.2506	0.3005

表4：抽取智能体在 n_{test}上的性能

反馈配置	***Bert* 精确率**	***Bert* 召回率**	***Bert F1***
自反馈	0.8681	0.8371	0.8523
Gemma 2 2B IT	**0.8918**	**0.8590**	**0.8751**
Gemma 3 1B IT	0.8664	0.8348	0.8502
Llama 3.2 3B Instruct	0.8665	0.8351	0.8505
Llama 3.2 1B Instruct	0.8661	0.8344	0.8499

表 5：反馈智能体对抽取质量的影响

LLM	$S_{\text{非结构化}}$	$S_{\text{结构化}}$
Llama 3.1 8B Instruct	55	145
Gemma 2 9B IT	**50**	**150**

表6：GPT-4 对结构化输入摘要的偏好（n_{sum}）

由经过指令微调的 Llama 3.2 3B Instruct 担任抽取智能体。我们比较了五种配置：（1）自我反馈（Madaan 等，2023），即使用经过指令微调的 Llama 3.2 3B Instruct 同时进行抽取和批评；（2）Llama 3.2 3B Instruct；（3）Llama 3.2 1B Instruct；（4）Gemma 2 2B IT；（5）Gemma 3 1B IT。后四种配置中的反馈智能体均为未经额外调优的预训练大语言模型。对于每种配置，我们对每份判决最多进行 10 轮优化，并采用第 4.1 节所述的耐心上限和容差阈值作为停止条件。表 5 报告了在同一测试集（n_{test}）上基于 BERT 的精确率、召回率和 F1。

5.3 摘要生成的影响

基于我们的抽取—反馈流程，我们评估了结构化事件时间线对法律摘要生成的下游益处。我们从测试集中随机抽取了200份判决（$n_{sum} = 200$），并使用两个开源大语言模型为每份判决生成两份摘要：Llama 3.1 8B Instruct（Grattafiori 等，2024）和 Gemma 2 9B IT（Google, 2024）。对于每份判决 J_j，我们生成了：

- 非结构化摘要：根据第4.2节中的公式10，以 J_j 为输入
- 结构化摘要：根据第 4.2 节中的公式 11，以 E_j 作为输入，其中 E_j 表示真实标注事件时间线。

为评估摘要质量，我们使用 GPT-4 作为自动评估器。对于每项判定，GPT-4 会获得 J_j、$S_{unstructured}$ 和 $S_{structured}$，并被要求选出最能提炼关键信息、保留法律语境且符合原始文档意图的摘要。

$$Best\ Summary = GPT\ 4(J_j, S_{unstructured}, S_{structured}) \tag{12}$$

A 结构化事件时间线显著提升了两种大语言模型架构的摘要生成质量，其中 Gemma 2 9B IT 的偏好度最高（75%）。这一强烈偏好表明，结构化时间线在不同基础模型（Meta 与 Google 架构）之间具有可迁移性和实用价值。我们的研究结果表明，抽取过程充当了知识预处理器，形成了高质量的信息瓶颈。通过向摘要生成模型提供简洁且按时间排序的法律叙事，我们使其能够专注于语言生成和高层次推理，而不必全面理解文档并筛选冗长的法律文本。

总而言之，抽取反馈评估验证了我们的双智能体框架在生成高质量事件时间线方面的有效性，而摘要生成实验则展示了结构化表示在法律人工智能任务中的下游优势。

6 数据与代码可用性

第 5.1 节中的代码、提示词、超参数、指令微调模型适配器以及完整的合成数据集均可在 GitHub 上获取（Tummepalli 2025）。所有资源均可公开访问。

7 对效度的威胁

尽管 LexChronos 展现出令人期待的成果，但某些因素可能会影响研究结果的普适性。(1) 合成数据集虽旨在模拟最高法院判决的结构与修辞特征，却并非源自真实案件。由于缺少真实司法复杂性，例如罕见的程序异常和不可预测的事实模式，其在现实场景中的适用性可能受到限制。(2) 该数据集仅涵盖最高法院分类框架中 48 个案件类别里的 25 个。尽管这些类别代表了案件量大且具有重要社会意义的领域，但排除

其他类别可能会引入主题偏差，并限制该框架对印度司法领域全貌的覆盖范围。(3) 当前评估仅限于英文判决。鉴于印度司法生态系统的语言多样性，以及地方语言材料的普遍存在，这一限制可能会影响该框架在多语言环境中的适应性。

8 结论与未来工作

本研究介绍并验证了*LexChronos*，这是一个用于从印度最高法院判决中抽取结构化事件时间线的智能体式人工智能框架。所提出的双智能体架构将指令微调抽取智能体与预训练反馈智能体相结合，为迭代式输出优化提供了一种系统性方法。为弥补该领域资源的不足，我们构建了一个包含 2,000 份标注判决的合成语料库，为印度法律事件抽取建立了基础基准。表现最佳的配置取得了 0.8751 的基于 BERT 的 F1 分数，展现出较高的准确率。此外，下游摘要生成实验证实了结构化时间线的实用价值。在大多数情况下，GPT-4 更偏好基于事件输入生成的摘要。这些发现凸显了超越传统非结构化方法、在法律人工智能应用中采用结构化表示的重要性。

展望未来，若干颇具前景的研究方向随之显现。关键的下一步是从合成数据转向真实世界数据集。尽管合成语料库支持了初步开发与评估，人工标注的最高法院数据集将验证该框架在实际场景中的稳健性，并成为法律人工智能社群的宝贵资源。此外，将 *LexChronos* 扩展至印度各级法院，包括高等法院和下级法庭，对于实现泛化能力至关重要。通过支持本土语言来应对语言多样性，将进一步提升其在印度司法生态系统中的适用性。

除了数据之外，该框架还可以扩展，以支持更高级的下游任务。虽然摘要生成曾作为最初的概念验证，但未来的工作将探索诸如判例映射等应用，使人工智能系统能够追溯判决的法律史及其影响；还包括论证生成和判决预测，这些任务需要对事件之间的逻辑关系和因果关系进行建模；以及知识库构建，利用时间关系将原始法律文本转化为相互连接的法律知识网络。这些方向标志着研究重点从基础信息提取转向面向任务的法律推理，使*LexChronos* 成为下一代法律人工智能系统的基础。

案件类别列表

表7列出了第3节所述、为数据集构建遴选的25种案件类别。

案件类别	
宪法	民法
银行与金融法	教育法
知识产权	税法
劳动与就业法	家庭法
消费者保护法	网络法
房地产与财产法	刑法
公务员服务法	合同法
环境法	公司法
行政法	保险法
卫生与医疗法	海事法
人权法	选举法
能源与矿业法	电信法
海关与消费税法	

表7：案件类别列表

参考文献

Achiam, J.; Adler,S.; Agarwal, S.; Ahmad,L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; 等。2023。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*。

Alan, R.; Lombardo, R.; Barbara, P.; 等. 2020. 将生物医学事件抽取作为序列标注。在*2020*年自然语言处理实证方法会议论文集（*EMNLP*），5357–5367。

Bj¨orne, J.; 和 Salakoski, T. 2018. 使用卷积神经网络和依存句法分析进行生物医学事件抽取。在*BioNLP 2018*研讨会论文集，98–108。

Chen, Y.; Zhou, T.; Li, S.; 和 Zhao, J. 2025年。一个用于文档级中文金融事件抽取的数据集。*Scientific Data*，12(1)：772。

Deeks, A.; 和 Hollis, D.2025年。大语言模型与国际法。《芝加哥国际法杂志》，26：117。

Filtz, E.; Navas-Loro, M.; Santos, C.; Polleres, A.; 和 Kir- rane, S. 2020。事件至关重要：从法院判决中抽取事件。在法律知识与信息系统，33–42。IOS Press。

Google。2024年。Gemma。

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan,A.; 等。2024年。Llama 3 模型家族。*arXiv* 预印本 *arXiv:2407.21783*。

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; 等。2025年。DeepSeek-R1：通过强化学习激发大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*。

Han, S.; Hao, X.; and Huang, H. 2018。基于在线中文新闻的商业分析事件抽取方法。《电子商务研究与应用》，28：244– 260。

Huang, J.; Gu, S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; and Han,J.2023. 大语言模型能够自我改进。载于

*2023*年自然语言处理实证方法会议论文集，1051–1068。Huang, K.-H.; and Peng, N. 2021. 通过高效的跨事件依赖端到端学习进行文档级事件抽取。*NAACL HLT2021*，36。Huang, Y.; Hu, N.; Li, K.; Wang, N.; and Lin, Z. 2024. 通过矩阵分块从原始文本中抽取金融事件。载于*2024*年计算语言学、语言资源与评估联合国际会议（*LREC-COLING 2024*）论文集，7035–7044。Hussain, A. S.; and Thomas, A. 2024. 司法实体抽取大语言模型：一项比较研究。*arXiv*预印本 *arXiv:2407.05786*。Indian Kanoon. 2025. Indian Kanoon. https://indiankanoon. org。访问日期：2025年9月29日。Kalamkar, P.; Agarwal, A.; Tiwari, A.; Gupta, S.; Karn, S.; and Raghavan, V. 2022. 印度法院判决中的命名实体识别。*arXiv*预印本 *arXiv:2211.03442*。Kaster, M.; Zhao, W.; and Eger, S. 2021. 通过沿语言因素解耦，探究基于BERT的评估指标的全局可解释性。载于*2021*年自然语言处理实证方法会议论文集，8912–8925。Liu, Y.; Fabbri, A. R.; Chen, J.; Zhao, Y.; Han, S.; Joty, S.; Liu, P.; Radev, D.; Wu, C.-S.; and Cohan, A. 2024. 基准评测大语言模型在指令可控摘要生成方面的生成与评估能力。载于计算语言学协会论文集：*NAACL 2024*，4481–4501。Liu, Y.; Feng, S.; Wang, D.; Zhang, Y.; and Sch¨utze, H. 2023. 评估无法评估的内容：生成式回复中不可评估的质量。*arXiv*预印本 *arXiv:2305.14658*。Lu, A.; Zhang, H.; Zhang, Y.; Wang, X.; and Yang, D. 2023. 通过提示约束界定大语言模型在开放式文本生成中的能力。载于计算语言学协会论文集：*EACL 2023*，1982–2008。Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. 自我精炼：通过自我反馈进行迭代式精炼。神经信息处理系统进展，36：46534–46594。Meta. 2024a. Llama 3.2 1B Instruct. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct. Hugging Face 模型库。Meta. 2024b. Llama 3.2 3B Instruct. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Hugging Face 模型库。Naik, V.; Patel, P.; and Kannan, R. 2023. 法律实体抽取：法律文档命名实体识别方法的实验研究。《国际高级计算机科学与应用期刊》，14(3)。新闻信息局、印度政府。2025. 人工智能在最高法院案件管理中的应用。https://www.pib.gov.in/PressReleasePage.aspx?PRID=2113224。访问日期：2025年9月29日。Rohit Singh. 2025年. 印度最高法院要求政府将人工智能工具整合进其案件管理系统。https://www.medianama.com/2025/07/223-sc-ai-tools-case-management-5cr-cases-pile-up-india/。访问日期：2025年9月29日。Shivananda, S.; Mathews, L. M.; Nikkam, S.; Kambli, P.; Ingale, R. D.; and Koparde, P. 2024年. 通过先进事件抽取革新法律案件分析。载于*2024*年科学与工程技术最新进展国际会议（*ICRASET*），1–6. IEEE. Sierra, G.; 等. 2018. 西班牙语法律文书中的事件抽取。载于第一届法律知识图谱语言资源与技术研讨会，36. 印度最高法院. 2025a. 案件类别。https://www.sci.gov.in/case-category/。访问日期：2025年9月29日。印度最高法院. 2025b. 新案件分类框架。https://cdnbbsr.s3waas.gov.in/s3ec0490f1f4972d133619a60c30f3559e/uploads/2025/05/2025051450.pdf。访问日期：2025年11月25日。Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ram´e, A.; Rivi`ere, M.; 等. 2025 年 . Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*。Team, Q. 2024年. Qwen2.5：一场基础模型的聚会。Tiwari, A.; Kalamkar, P.; Banerjee, A.; Karn, S.; Hemachandran, V.; and Gupta, S. 2024年. Aalap：面向印度法律及辅助法律职能的人工智能助手。*arXiv*预印本 *arXiv:2402.01758*。Tummepalli, A. C. 2025年. LexChronos代码库。https://github.com/chandrahas316/LexChronos。访问日期：2025年12月15日。Wang, N.; Peng, Z.; Que, H.; Liu, J.; Zhou, W.; Wu, Y.; Guo, H.; Gan, R.; Ni, Z.; Yang, J.; 等. 2024年. RoleLLM：对大语言模型角色扮演能力的基准评测、激发与增强。载于计算语言学协会*ACL 2024 Findings*，14743–14777。Xian, G.; Du, S.; Tang, X.; Shi, Y.; Jia, B.; Tang, B.; Leng, Z.; and Li, L. 2024年. DLEE：面向中文文档级法律事件抽取的数据集。《神经计算与应用》，36(25)：15581–15597。Xu, H.; Peng, B.; Awadalla, H.; Chen, D.; Chen, Y.-C.; Gao, M.; Kim, Y. J.; Li, Y.; Ren, L.; Shen, Y.; 等. 2025年. Phi-4-mini-reasoning：探索小型推理语言模型在数学中的极限。*arXiv*预印本 *arXiv:2504.21233*。Xu, N. 2024年. 基于机器学习的智能经济法律系统中的事件实体抽取研究。载于*2024*年图像处理、智能控制与计算机工程国际会议论文集，164–168。Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; 等. 2025年. Qwen3技术报告。*arXiv*预印本 *arXiv:2505.09388*。Yue, L.; Liu, Q.; Zhao, L.; Wang, L.; Gao, W.; and An, Y. 2024年. 事件驱动的刑事法庭观点生成

与协作式（大）语言模型。载于第*47*届国际*ACM SIGIR*信息检索研究与发展会议论文集，2221–2230。
Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019年. BERTScore：使用BERT评估文本生成。*arXiv*预印本 *arXiv:1904.09675*。Zhao, B.; Zhao, Y.; and Mao, Y. 2024年. 一种基于事件抽取的司法案例知识图谱构建方法。载于*2024*年第九届智能信息技术国际会议论文集，62–69。