Enhancing Zero-Shot Temporal Entailment in Legal Documents via Hybrid Element Fact Extraction

Gwang-Jae Won[1], and Ro-Seop Park[1*]

[1] *Hallym University, 1 Hallymdaehak-gil, Chuncheon-si, Gangwon-do 24252, Republic of Korea*

Abstract

Zero-shot temporal reasoning over legal documents is complicated by the frequent omission of explicit timestamps and the abundance of textual noise, requiring models to infer relative sequences rather than merely extract dates. To address this, we introduce an Element Fact extraction pipeline that decomposes legal narratives into structured, filtered representations through phase-based classification and systematic admissibility criteria. Evaluated on the LexTime dataset(514 instances) across multiple LLMs, a hybrid context engineering comprising both the original paragraph and the extracted Element Facts yielded an accuracy of 75.3% and a contradiction detection F1 score of 74.4% with GPT-OSS-120B. Furthermore, an ablation study indicates that these structured facts function as an attentional signal amplifier. Supplied alongside the raw text, they direct the LLM's attention to core conflicting points, mitigating both the tendency for false positive entailments induced by unfiltered narratives and the propensity for false negative contradictions caused by isolated facts. These results provide evidence that structured logic and natural narrative play synergistic and complementary roles in legal temporal reasoning.

Keywords

Zero-shot Temporal Reasoning, Legal Natural Language Processing, Element Fact Extraction, Context Engineering, Large Language Models

1. Introduction

In criminal cases, the rapidly expanding volume of unstructured digital evidence frequently induces cognitive overload and confirmation bias, impeding the discovery of substantive truth [1]. While automated timeline reconstruction using large language models (LLMs) has emerged as a crucial mechanism to mitigate this information overload and support case analysis [2]–[4], the direct application of these models to legal documents presents critical structural challenges.

First, temporal reasoning in the legal domain is complicated by the pervasive omission of e[1]xact timestamps. Key events frequently lack precise temporal markers, requiring models to infer relative chronological orders and causal relationships rather than merely extract explicit dates. This domain-specific characteristic makes zero-shot temporal entailment fundamentally a reasoning task. Second, legal texts are densely populated with extraneous noise, including procedural metadata, formulaic recitals, and speculative passages. Current single-pass LLM approaches, which process the entire raw text simultaneously, often falter under this high noise ratio. They easily overlook subtle temporal contradictions or generate contextually plausible yet factually incorrect temporal associations—a phenomenon commonly referred to as hallucination [5].

To address these limitations, this paper proposes an Element Fact extraction pipeline that decomposes the narrative structure of a legal document into a systematic, noise-attenuated

[1]*RELAF'26: Reasoning with Evidence in Law Enforcement and Forensics, Workshop at ICAIL 2026, Singapore.*

[*] Corresponding author.

rhkdwo875@gmail.com (G. J. Won); rspark@hallym.ac.kr (R. S. Park)

0009-0008-1006-2623 (G. J. Won); 0009-0003-4957-7804(R. S. Park)

representation. The pipeline operates in three concise stages: (1) pre-filtering non-analytical sentences; (2) classifying sentences into narrative phases; and (3) converting sentences that satisfy admissibility criteria into structured Element Facts by integrating refined factual summaries with explicit temporal metadata. Evaluated on the LexTime dataset [6], a hybrid context configuration—supplying both the original paragraph and the extracted Element Facts—demonstrates the highest aggregate performance, indicating that structured logic and natural narrative serve synergistic and complementary roles in legal temporal reasoning.

2. Related Work

2.1. Multi-Dimensional Temporality of Legal Cases (LexTime)

Legal texts contain dense explicit temporal markers, yet authentic legal temporal reasoning fundamentally requires the capacity to infer implicit events logically entailed by the surrounding context. Because the LexTime dataset [6] is specifically formulated to evaluate sequence inference among such implicit legal events in the absence of explicit temporal anchors, it serves as the optimal benchmark for validating our context-driven methodology. While various approaches integrate symbolic temporal reasoning with machine learning for event segmentation and timeline summarization [7]–[11], these methods significantly degrade when applied to legal documents, where causal relations are interleaved with dense extraneous noise [12]. Furthermore, general-domain annotation standards focus primarily on surface-level actions [13], lacking mechanisms to separate substantive facts from rhetorical content.

2.2. Generative LLMs and Cognitive Scaffolding

Information extraction has transitioned toward generative paradigms utilizing instruction-tuned LLMs [14], [15], dynamic guidelines [16], and multi-stage prompting [17]. However, LLMs frequently generate unconstrained inferences (hallucinations) over complex event topologies [18]. To mitigate this, event extraction requires a "cognitive scaffold"—a structural interface providing explicit schemas to ground LLM outputs [15]. Our Element Fact pipeline addresses this limitation by filtering texts prior to model ingestion and enforcing strict admissibility criteria, thereby anchoring verifiable claims in the source text.

2.3. Computational Bottlenecks and Explainability

Processing voluminous legal records frequently exceeds the working memory of contemporary models. Although approaches like storyline visualizations, graph transformers, and joint extraction models address long-context challenges [19]–[22], extraction inaccuracies tend to propagate cumulatively through downstream relational components [23], [24]. Furthermore, relying entirely on opaque deep-learning models in critical judicial domains presents inherent limitations [27]. While interactive systems [25] and contestable AI [26] offer human oversight, the underlying extraction models remain black boxes. The proposed representation mitigates these bottlenecks by reducing dimensionality and establishes an auditable framework by grounding extracted facts in specific textual spans.

3. Methodology

3.1. Overview of the Element Fact Pipeline

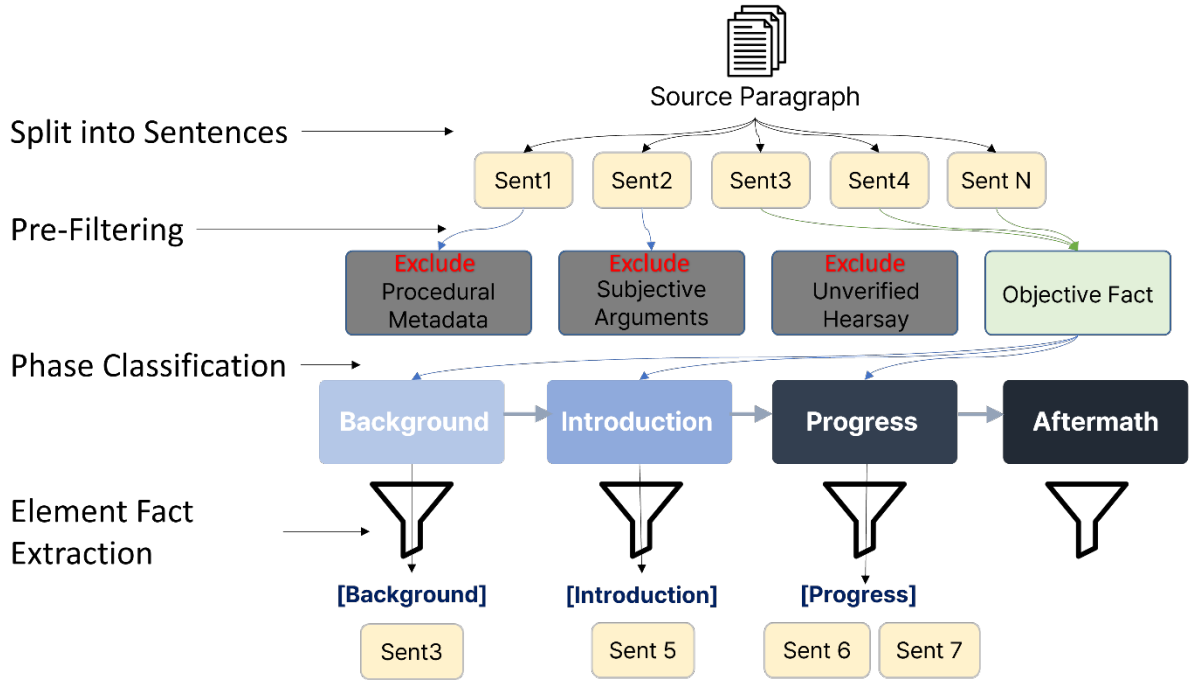

Fig. 1. Overall pipeline of the proposed Element Fact extraction system

To address the contextual noise limitations previously identified, this study proposes an Element Fact extraction pipeline that decomposes a legal document into a structured, noise-attenuated representation suitable for temporal reasoning. The pipeline reorganizes the narrative structure of the source text into a form optimized for temporal and causal analysis, operating in three stages detailed in the subsequent subsections.

3.2. Sentence Processing and Pre-Filtering

In the first stage, the source paragraph is segmented into sentences utilizing a custom heuristic regular expression parser developed in Python. Subsequently, LLM(GPT-4o) is employed to categorize each sentence based on its relevance to temporal reasoning. During this process, the language model systematically annotates sentences that consist exclusively of procedural information(such as statute citations, case numbers, court venues, or formulaic recitals), speculative or hedged passages, and other non-analytical material are tagged for downstream filtering. The objective at this stage is not to permanently eliminate such sentences, but to annotate them so as to ensure that subsequent stages can apply filtering in a principled and reversible manner.

Table 1. Annotation guidelines for filtering sentence

Category	Filtering Decision	Description	Example
Substantive Fact	Pass (Retained)	Objective events, actions, state transitions, or causality that form the core chronological narrative of the legal case. These sentences serve as the foundation for temporal reasoning.	"Trachte received the award on december 1, 2023, and the amended award on december 11..."
Procedural Metadata	Exclude (Filtered)	Mechanical boilerplate text, document headers, filing dates, venue statements, and case numbers that do not contribute to the factual narrative.	"case 3:23-cv-00466 document 1 filed 04/13/23 page 7 of 10"
Subjective Arguments	Exclude (Filtered)	Evaluative claims, legal conclusions, or opinions asserted by a party without establishing a concrete chronological event.	"and the conduct here is especially egregious because, under the iam constitution, she has a special obligation..."
Hearsay	Exclude (Filtered)	Second-hand information, rumors, or claims prefaced by legal boilerplate indicating a lack of direct evidentiary backing at the time of writing.	"upon information and belief, the union authorized the union members concerted action of storming meetings..."

3.3. Phase-Based Classification

In the second stage, the tagged sentences are classified according to their position within the narrative flow of the case. Four phases are distinguished, corresponding to the principal stages through which a typical legal narrative unfolds.

(1) Background: pre-event circumstances and the relationship between the parties prior to the incident.

(2) Introduction: the inception of the event, its precipitating act, and the immediate initial response.
(3) Progress: the principal acts and factual relations that constitute the body of the event, together with any subsequent developments.

(4) Aftermath: the consequences of the event, including damage, the filing of a complaint, and any subsequent settlement or remedial measures.

In practice, this phase-based classification is implemented automatically via zero-shot prompting of a Large Language Model (LLM). It explicates the temporal flow of the case and provides a macro-level temporal anchor for the subsequent element fact extraction.

Table 2. Classification of examples for the four phases

Phase	Example
Background	despite no change in audreys need for daily skilled nursing services and no change in the language of the plan.
Introduction	After January 22, 2021, BCBSIL began denying Mr. Lou's claims for Audrey's care.
Progress	On February 16, 2021, BCBSIL denied Mr. Lou's request for benefits for the period from January 28, 2021 through April 30.
Aftermath	On April 13, 2021, Maxim Healthcare Services submitted an administrative appeal of BCBSIL's claim denial on Mr. Lou's behalf.

3.4. Extract Element Fact

Conceptually, an Element Fact within this framework extends beyond the traditional Natural Language Processing (NLP) definition of an event "trigger," which typically isolates surface-level keywords or verbs. Instead, an Element Fact is defined as the atomic unit of factual transition that carries legal significance. Formally, it operates as a complete propositional structure that semantically binds a triggering action with its essential arguments, including actors, objects, timestamps, and the resulting qualitative transition of the legal state. By encapsulating these dynamics into indivisible units, Element Facts function as "cognitive anchors" that bridge raw narratives and legal evaluations, rather than acting solely as isolated data points.

In the third stage, a structured informational unit, designated as an Element Fact, is extracted from sentences classified into one of the four aforementioned phases. Information categorized as non-analytical during the initial stage, including identification metadata, procedural details devoid of substantive events, and subjective sentiment lacking associated act, is systematically excluded from this extraction process.

To ensure that the resulting Element Facts capture substantive qualitative transitions of legal states rather than extraneous details, the extraction process is governed by five admissibility criteria. These criteria are theoretically grounded in the formal structural components of a legal narrative "Trigger" (Timestamp, Pre/Post-State, Triggering Event, Causal Chain, and Legal Function).

A sentence is retained only if it satisfies the following conditions:

(1) Temporal Specificity: The temporal location of the act can be identified, either explicitly or through clear contextual inference. (2) State Change: The act produces a physical, legal, or perceptual change of state for the involved parties. (3) Causal Importance: The act is constitutive of the narrative structure rather than being incidental. (4) Legal Relevance: The act bears directly or indirectly on a doctrinal element, such as deception, disposition, or violence. (5) Factual Basis: The act is described as an objective fact rather than a subjective emotional report.

By embedding these criteria as constraints within the LLM prompt, the output of this stage yields a list of Element Facts that constitute a noise-attenuated, temporally and causally organized representation of the case. This structure facilitates direct processing by a downstream LLM.

4. Experimental Design

4.1. Dataset

The empirical evaluation is conducted utilizing the LexTime dataset [6], which comprises 514 instances specifically curated to assess temporal event ordering capabilities within the legal domain. Each instance encompasses a source paragraph extracted from U.S. Federal Complaints, paired with a temporal reasoning query designed to determine the temporal relationship between two specified events. The classification task is strictly binary; each query is designated as either yes (indicating logical consistency with the source)(n=256) or no (denoting inconsistency)(n=255).

Table 3. Comparison of input context representations between element fact and Lextime[6]

	Element Fact	Lextime
Context (paragraph)	[1] (Introduction) PBGC requested Safety Light to provide documents and information. | Time: March 9, 2022 [2] (Progress) PBGC sent the information request letter via electronic mail to Silverthorn. | Time: [3] (Progress) On April 12, 2022, PBGC issued a subpoena to Safety Light requiring document production by May 3, 2022. | Time: April 12, 2022 [4] (Progress) Respondents failed to respond to the Safety Light subpoena within the prescribed time. | Time: [5] (Progress) On May 4, 2022, Silverthorn requested an extension to respond until the end of May. | Time: May 4, 2022 [6] (Progress) PBGC agreed to the extension with documents to be provided on a rolling basis. | Time:	as part of its investigation, pbgc requested by letter dated march 9, 2022, that safety light provide to pbgc certain documents and information relating to the plan and the financial condition of safety light [...] on april 12, 2022, pbgc issued a subpoena to safety light [...] requiring production [...] on or before may 3, 2022 [...] however, on may 4, 2022, [...] silverthorn requested an extension to respond until the end of may. pbgc agreed to the extension [...] respondents did not meet this new deadline nor communicate with pbgc in the interim.
Query	Silverthorn's request for an extension follows PBGC's follow-up email regarding the Safety Light subpoena.	
Answer	entailment	

4.2. Compared Methods

Three context engineerings are compared, each evaluated utilizing an identical model and prompt architecture(Appendices Table 6-7). (1) Baseline(Paragraph Only): The unstructured source paragraph is supplied to the LLM without preprocessing, representing the conventional methodology. (2) Element Fact Only: The original paragraph is omitted, and solely the structured array of Element Facts produced by the pipeline is supplied as contextual input. (3) Hybrid (Paragraph and Element Fact): Both the source paragraph and the Element Facts are supplied as context, enabling the model to simultaneously process the natural language narrative and the noise-attenuated structured representation.

4.3. Evaluation Metrics

Performance is quantified utilizing overall accuracy and the macro F1 score across the binary entailment classifications. Furthermore, to facilitate a granular analysis of the models' inherent inferential biases across different context engineerings, class-wise precision, recall, and F1 scores are additionally reported for both the entailment and contradiction labels. This approach systematically distinguishes methods exhibiting a propensity for false positive entailments from those demonstrating a predisposition toward contradictions.

4.4. Experimental Setup

The evaluations employ GPT-4o as the primary reasoning model within a zero-shot framework. The original LexTime study [6] benchmarked a broad spectrum of models, spanning proprietary LLMs (GPT-4o, GPT-4 Turbo), large open-source LLMs (Mistral-123B, LLaMA 3.1-70B), smaller LLaMA variants (3.1-8B, 3.2-3B, 3.2-1B in both instruct and base forms), and a lighter-weight encoder-decoder baseline (Flan-T5-780M). Reported few-shot accuracies on the full dataset ranged from 48.7% (LLaMA 3.2-1B Base) to 77.4% (GPT-4o), with Flan-T5-780M at 55.5% and Mistral-123B at 73.9%. GPT-4o was selected for our study as it achieved the highest reported accuracy on the full dataset in the few-shot setting (77.4%), thereby serving as a robust upper-bound reference for legal temporal reasoning capabilities [6]. Note that our study is conducted under a zero-shot framework to isolate the effect of context engineering; the paragraph-only Baseline reported in Section 5 (71.7% for GPT-4o) is therefore naturally lower than the few-shot reference (77.4%), and provides the direct comparison point for our Element Fact configurations.

5. Results and Analysis

Table 4. Aggregate performance and detailed metrics of the three context configurations across multiple LLMs on the LexTime dataset. Reported figures are means over three independent runs, with all values reported in percentages. Bold values indicate the best score in the Average row and the highest scores across all models.

Model	Method	Acc	Macro F1	Ent. P	Ent. R	Ent. F1	Cont. P	Cont. R	Cont. F1
GPT-4o	Baseline	71.7	71.4	68.1	**82.5**	**74.6**	**77.4**	60.8	68.1
	EF Only	68.4	68.2	71.5	61.9	66.4	65.9	**74.9**	70.1
	Hybrid	**73.1**	**73.0**	72.0	76.2	74.0	74.3	69.9	**72.0**
GPT-OSS-120B	Baseline	72.7	72.6	70.0	**81.6**	75.4	77.3	63.7	69.8
	EF Only	68.2	68.3	72.5	60.6	66.0	65.9	**75.8**	70.5
	Hybrid	**75.3**	**75.5**	73.9	79.3	**76.5**	**78.0**	71.2	**74.4**
Gemma-4-31B-it	Baseline	**53.4**	60.4	**74.9**	**65.5**	**69.9**	66.3	41.2	50.8
	EF Only	48.2	55.7	72.8	45.4	55.9	60.6	51.0	55.4
	Hybrid	50.2	**60.0**	73.8	43.2	54.5	**76.5**	**57.3**	**65.5**
Average	Baseline	65.9	68.1	71.0	**76.5**	**73.3**	73.7	55.2	62.9
	EF Only	61.6	64.1	72.3	56.0	62.8	64.1	**67.2**	65.3
	Hybrid	**66.2**	**69.5**	**73.2**	66.2	68.3	**76.3**	66.1	**70.7**

5.1 Aggregate Performance Across Diverse LLMs

Table 4 reports the aggregate and class-wise performance across three different LLMs, indicating that the effectiveness of the proposed framework generalizes across model architectures, particularly for contradiction detection. On average, the Hybrid configuration achieves the highest overall performance with an accuracy of 66.2% and a Macro F1 of 69.5%. This configuration modulates the trade-offs observed in the single-context baselines and reaches the highest average Contradiction F1 at 70.7%. Furthermore, when supplied with the Hybrid context, the open-source GPT-OSS-120B model surpassed the proprietary GPT-4o baseline, reaching the maximum observed performance with an accuracy of 75.3% and a Macro F1 of 75.5%. These results suggest that structured factual scaffolding enables open-source architectures to perform rigorous legal temporal reasoning.

A notable exception is Gemma-4-31B-it, where the Hybrid configuration underperforms the Baseline in overall accuracy (50.2% vs. 53.4%), suggesting that smaller-capacity models may struggle to jointly integrate raw narrative and structured facts within a single context window. Nevertheless, even in this case the Hybrid setting substantially improved Contradiction F1 by 14.7 percentage points over the Baseline, indicating that the attentional amplification effect on contradiction detection persists across model scales, while the overall accuracy gain is more sensitive to model capacity.

5.2. Trade-Off Between the Two Single-Context Conditions

An analysis of the averaged Precision and Recall metrics reveals a consistent trade-off that indicates distinct behavioral tendencies inherent to LLMs in single-context conditions. The Baseline exhibits a propensity for false positive entailments, which can be characterized as a gullible bias. When processing the natural narrative flow of the original text, the models frequently default to Entailment predictions, resulting in an average Entailment Recall of 76.5%. However, the presence of extensive textual noise impedes the detection of subtle temporal discrepancies, which reduces the average Contradiction Recall to 55.2%.

Conversely, the Element-Fact-only condition demonstrates a predisposition toward contradictions, characterizable as an overly suspicious bias. Without rhetorical noise, the models identify conflicting facts, increasing the average Contradiction Recall to 67.2%. Deprived of contextual connective tissue, however, the models show reduced capacity to recognize implicitly entailed events, which lowers the average Entailment Recall to 56.0% and decreases overall accuracy.

5.3. Synergy in the Hybrid Condition Across Models

The Hybrid configuration mitigates both extremes by integrating their strengths, an effect most pronounced in higher-capacity models (GPT-4o, GPT-OSS-120B) and observable, particularly in contradiction detection, across all evaluated models. Supplying both representations allows the natural narrative to provide the contextual links necessary for valid entailments. Simultaneously, the structured Element Facts function as explicit referential constraints that amplify attentional signals. This integration facilitates the capture of contradictions that the Baseline omitted, increasing the average Contradiction Precision to 76.3% and achieving the highest average Contradiction F1 of 70.7%. For example, in the Gemma-4-31B-it model, the Hybrid context increased Contradiction Precision to 76.5% and Recall to 57.3%, raising its Contradiction F1 by 14.7 percentage points over the Baseline. These empirical gains suggest that structured logic and natural narrative serve complementary roles in legal reasoning.

5.4. Discussion

5.4.1. Mechanism of Synergy: Element Facts as an Attentional Signal Amplifier

To delineate the underlying mechanism driving the Hybrid model's performance, we conducted an exploratory ablation study on a subset of 20 challenging instances to qualitatively analyze the impact of individual admissibility criteria. Given the limited sample size, the component-level conclusions presented in this subsection should be treated as illustrative rather than conclusive, and are intended to generate hypotheses about the functional roles of individual criteria rather than to establish their relative importance definitively.

Table 5. Exploratory Ablation Study (N=20) isolating the impact of admissibility criteria.

Condition	Macro F1	Entailment F1	Contradiction F1
Control (All Criteria Applied)	80.00%	80.00%	80.00%
State Change	80.0%	80.00%	80.0%
Causal Importance	79.80%	81.80%	77.80%
Factual Basis	74.90%	76.20%	73.70%
Legal Relevance	74.40%	78.30%	70.60%
Temporal Specificity	**69.70%**	**72.70%**	**66.70%**

The results suggest two distinct functional roles within the criteria.

First, Temporal Specificity appears to act as the primary reference axis for comparison in this exploratory sample. Its removal causes the most substantial performance degradation(Contradiction F1 dropping by 13.3pp), which is consistent with the interpretation that without an explicit or inferred temporal baseline, the relative deductive reasoning structure is impaired.

Second, Legal Relevance and Factual Basis appear to contribute to noise reduction, potentially preventing subjective evaluations and extraneous metadata from obscuring the structured timeline. These observations offer a plausible interpretation of why the Hybrid model achieves optimal performance.

In the Hybrid setup, the raw paragraph is still provided to LLM, meaning the textual noise is not physically deleted from the context window. Instead, the rigorously filtered and temporally aligned Element Facts act as an "Attentional Signal Amplifier" by structuring the contextual input. They direct the model’s computational focus directly onto the core conflicting points within the dense textual data, effectively suppressing the propensity for false positive entailments and enabling highly rigorous contradiction detection without losing the necessary narrative context.

5.4.2. Broader Implications for Legal AI

Beyond empirical performance gains, this mechanism directly addresses the real-world challenges outlined in the Introduction. In complex legal investigations, human professionals are highly

vulnerable to cognitive overload and confirmation bias, often missing critical contradictions buried within voluminous evidence. The Baseline's susceptibility to false positives demonstrates that generative LLMs are susceptible to a similar form of automated confirmation bias when overwhelmed by textual noise.

By utilizing the Element Fact pipeline as a cognitive scaffold, the Hybrid approach empirically demonstrates that structured logic and natural narrative are highly synergistic. Furthermore, because each extracted fact is explicitly grounded in stringent admissibility criteria and linked to specific narrative phases, this architecture inherently enhances explainability. It allows human investigators to trace the model's temporal inferences back to verified textual spans, mitigating the opacity limitations of generative AI and establishing a more robust, auditable framework for judicial applications.

5.5. Limitations

Several limitations should be considered when interpreting the findings of this study. These span issues of generalization scope, methodological independence, pipeline transparency, and statistical robustness.

The most fundamental concerns the scope of generalization. Although our motivation stems from criminal investigation document analysis, the LexTime benchmark consists exclusively of U.S. Federal civil complaints. The noise categories targeted by our pipeline, such as statute citations, formulaic recitals, and procedural metadata, are characteristic of formal civil filings. Whether both the filtering taxonomy and the four-phase schema transfer to criminal investigative documents remains an open empirical question, and cross-domain validation on criminal case records is therefore a necessary next step.

A related concern is methodological independence. GPT-4o serves both as the extraction model in the pipeline and as one of the downstream reasoning models, which constitutes a potential confound with respect to model self-consistency effects. Mitigating this, the strongest Hybrid gains were observed with GPT-OSS-120B, a distinct model with different weights and training. Gemma-4-31B-it, which comes from an entirely different model family, likewise exhibited substantial improvement in contradiction detection under the Hybrid setting. This cross-architecture consistency suggests that the observed synergy is not solely attributable to extractor–reasoner self-consistency. Nonetheless, GPT-4o and GPT-OSS-120B share a common developer, and replicating the extraction stage with structurally unrelated models such as Claude or Llama remains an important direction for future work.

Beyond these external and methodological considerations, the pipeline's internal reliability also warrants further scrutiny. Because our system relies on LLM-based processing at every stage, including pre-filtering, phase classification, and fact extraction, errors introduced at earlier stages may propagate invisibly to downstream reasoning. This study does not report per-stage quality estimates, and quantifying stage-wise extraction quality through human-annotated samples is a priority for future work.

Finally, our exploratory ablation experiment (N=20) provides only preliminary evidence for the functional roles of individual admissibility criteria. While this analysis motivates our interpretation of Element Facts as an attentional signal amplifier, the sample size inherently limits the generalizability of component-level findings. In particular, the observed ordering of criteria by impact, most notably the apparent primacy of Temporal Specificity, should be regarded as a

hypothesis to be tested rather than a definitive ranking. Verifying the statistical significance and precise contribution of each of the four narrative phases and the five admissibility criteria across a comprehensive, large-scale dataset remains a necessary task for future work.

6. Conclusion

This study presents an Element Fact extraction pipeline designed to decompose legal narratives into structured, noise-filtered representations through phase-based classification and predefined admissibility criteria.

Evaluations on the LexTime dataset, across multiple large language models of varying capacities indicate that the proposed Hybrid context engineering consistently improved contradiction detection across all evaluated models, while overall accuracy gains were most pronounced in higher-capacity models such as GPT-4o and GPT-OSS-120B. Notably, the Hybrid approach achieved the highest performance with GPT-OSS-120B, reaching an accuracy of 75.3% and a contradiction detection F1 score of 74.4%.

Mechanism analysis demonstrates that these structured facts function as an attentional signal amplifier, guiding the LLM's focus toward core conflicting points and mitigating inherent inferential biases, with the strength of this effect varying by model capacity.

Building upon the exploratory ablation analysis, future research directions include conducting large-scale statistical validations of individual pipeline components, extending the framework to additional legal subdomains, examining its behavior under varying capacities of open-source models, and developing automatic audit procedures that exploit the span-level grounding of the extracted facts.

Acknowledgements

This research was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) funded by the Ministry of Science and ICT(MSIT)(No. RS-2025-02318092)

Declaration on Generative AI

During the preparation of this work, the author used Google Gemini in order to translate parts of the manuscript from Korean to English and to improve the language and readability. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] F. J. Bex, P. J. van Koppen, H. Prakken, and B. Verheij, "A hybrid formal theory of arguments, stories and criminal evidence," *Artificial Intelligence and La*w, vol. 18, no. 2, pp. 123–152, 2010.

[2] H. Surden, "Machine learning and law," *Washington Law Revie*w, vol. 89, no. 1, pp. 87–115, 2014.

[3] Won, G. J., Park, R. S., Park, S. M., "an AI-based Hierarchical Timeline Generation and Verification System to Mitigate Confirmation Bias in Criminal Investigation," *Korean Police Studies Review*, vol. 23, no. 3, pp. 89-126, 2025.

[4] S. W. van den Braak, "Sensemaking software for crime analysis," Ph.D. dissertation, Utrecht University, 2010.

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[6] C. Barale, L. Barrett, V. S. Bajaj, and M. Rovatsos, "LEXTIME: A benchmark for temporal ordering of legal events," *arXiv preprint arXiv:2506.04041*, 2025.

[7] A. Leeuwenberg and M.-F. Moens, "A Survey on Temporal Reasoning for Temporal Information Extraction from Text," *Journal of Artificial Intelligence Research*, vol. 66, pp. 341-380, 2019.

[8] S. Michelmann, M. Kumar, K. A. Norman, and M. Toneva, "Large language models can segment narrative events similarly to humans," *arXiv preprint arXiv:2301.10297*, 2023.

[9] M. R. Qorib, Q. Hu, and H. T. Ng, "Just What You Desire: Constrained Timeline Summarization with Self-Reflection for Enhanced Relevance," *arXiv preprint* arXiv:2412.17408, Dec. 2024.

[10] K. T. Ong et al., "Towards Lifelong Dialogue Agents via Timeline-based Memory Management," in Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 8631–8645.

[11] S. Li, R. Zhao, M. Li, H. Ji, C. Callison-Burch, and J. Han, "Open-domain hierarchical event schema induction by incremental prompting and verification," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023.

[12] F. A. Tan, H. Hettiarachchi, A. Hürriyetoğlu, T. Caselli, O. Uca, F. F. Liza, and N. Oostdijk, "Event Causality Identification with Causal News Corpus - Shared Task 3, CASE 2022," in Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), Dec. 2022, pp. 195–208.

[13] J. Aguilar et al., "A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards," *in Proc. EVENTS Workshop*, 2014.

[14] É. Simon et al., "Generative Approaches to Event Extraction: Survey and Outlook," *in Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pp. 73–86, 2024

[15] B. Li et al., "Event Extraction in Large Language Model: A Holistic Survey of Method, Modality, and Future," *arXiv preprint arXiv:2512.19537*, 2025.

[16] S. Srivastava, S. Pati, and Z. Yao, "Instruction-Tuning LLMs for Event Extraction with Annotation Guidelines", *arXiv:2502.16377v1*, 2025.

[17] E. Cai and B. O'Connor, "A Monte Carlo Language Model Pipeline for Zero-Shot Sociopolitical Event Extraction," *in Proc. 61st Annu. Meeting Assoc. Comput. Linguist.*, pp. 1651–1665, 2023.

[18] K. H. Huang et al., "TEXTEE: Benchmark, Reevaluation, Reflections, and Future Challenges in Event Extraction," *in Findings of the Association for Computational Linguistics: ACL 2024, Aug*. pp. 12804–12825, 2024.

[19] Y. Tanahashi and K.-L. Ma, "Design considerations for optimizing storyline visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2679–2688, 2012.

[20] Y. Zeng, X. Jin, S. Guan, J. Guo, and X. Cheng, "Event coreference resolution with their paraphrases and argument-aware embeddings," *in Proceedings of the 28th International Conference on Computational Linguistics (COLING),* pp. 3084–3094, 2020

[21] R. Chaturvedi et al., "Temporal Relation Extraction in Clinical Texts," *Proceedings of the 63rd ACL*, pp. 25765–25788, 2025

[22] J. Araki and T. Mitamura, "Joint Event Trigger Identification and Event Coreference Resolution with Structured Perceptron," *in Proc. 2015 Conf. Empirical Methods in Natural Language Processing (EMNLP),* Lisbon, Portugal, pp. 2074–2080, 2015

[23] S. Guan, X. Cheng, L. Bai, F. Zhang, Z. Li, Y. Zeng, X. Jin, and J. Guo, "What is Event Knowledge Graph: A Survey," *arXiv preprint arXiv:2112.15280*, 2021.

[24] H. Zhang, M. Chen, H. Wang, Y. Song, and D. Roth, "Analogous Process Structure Induction for Sub-event Sequence Prediction," *in Proc. EMNLP*, pp. 1541–1550, 2020

[25] N. Lagos, F. Segond, S. Castellani, and J. O'Neill, "Event Extraction for Legal Case Building and Reasoning," *in Intelligent Information Processing V, Z. Shi, S. Vadera, A. Aamodt, and D. Leake, Eds. Berlin, Heidelberg: Springer*, pp. 92–101, 2010

[26] F. Maoro and M. Geierhos, "Contestable AI for criminal intelligence analysis: improving decision-making through semantic modeling and human oversight," *Front. Artif. Intell.*, vol. 8, art. no. 1602998, doi: 10.3389/frai.2025.1602998, 2025

[27] X. Zhao et al., "A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers," *ACM Comput. Surv.*, vol. 56, no. 11, Art. no. 293, doi: 10.1145/3674501, 2024.

Appendices

Table 6. Prompt templates for evaluation, on three types of prompts used: Lextime(Paragraph only), Element Fact(EF only), Hybrid(Paragraph + EF).

Method	Prompt
Lextime	Given the context, the task is to answer the query with “yes” or “no. context: “{Paragraph}” query: “{query}” answer:
EF Only	Given the context, the task is to answer the query with “yes” or “no. context: “{Element Fact}” query: “{query}” answer:
Hybrid (EF+Paragraph)	Given the context, the task is to answer the query with “yes” or “no. context: “{Paragraph \n Element Fact}” query: “{query}” answer:

Table 7. Examples by 3-Methodology.

Method	Examples
Lextime	Given the context, the task is to answer the query with “yes” or “no. context: “he new policy, which was facially discriminatory based on age, was approved by the defendants on the executive counsel on or shortly after june 17, at the time the new policy was approved by the executive council, all of the defendants on the executive council planned on retiring at age 65 and, as a result, knew that pantoja was the only member of the executive council that would be adversely affected by the policy. the new policy was aimed at removing pantoja because he had exercised his right to free speech under the lmrda. e. pantojas removal from elected office on april 1, 2022 pursuant to the mandatory retirement policy.” query: “Pantoja's removal from elected office follows the approval of the new policy by the defendants.” answer: yes
EF Only	Given the context, the task is to answer the query with “yes” or “no. context: “[1] (Background) The new policy, discriminatory based on age, was approved by the defendants on the executive council. | Time: shortly after June 17 [2] (Background) All defendants planned on retiring at age 65 and knew

	Pantoja would be adversely affected by the policy. | Time: shortly after June 17 [3] (Introduction) The new policy was aimed at removing Pantoja for exercising his right to free speech under the LMRDA. | Time: none [4] (Aftermath) Pantoja's removal from elected office on April 1, 2022, pursuant to the mandatory retirement policy. | Time: April 1, 2022" query: "Pantoja's removal from elected office follows the approval of the new policy by the defendants." answer: yes
Hybrid (EF+Paragraph)	Given the context, the task is to answer the query with "yes" or "no. context: ""he new policy, which was facially discriminatory based on age, was approved by the defendants on the executive counsel on or shortly after june 17, at the time the new policy was approved by the executive council, all of the defendants on the executive council planned on retiring at age 65 and, as a result, knew that pantoja was the only member of the executive council that would be adversely affected by the policy. the new policy was aimed at removing pantoja because he had exercised his right to free speech under the lmrda. e. pantojas removal from elected office on april 1, 2022 pursuant to the mandatory retirement policy. [1] (Background) The new policy, discriminatory based on age, was approved by the defendants on the executive council. | Time: shortly after June 17 [2] (Background) All defendants planned on retiring at age 65 and knew Pantoja would be adversely affected by the policy. | Time: shortly after June 17 [3] (Introduction) The new policy was aimed at removing Pantoja for exercising his right to free speech under the LMRDA. | Time: none [4] (Aftermath) Pantoja's removal from elected office on April 1, 2022, pursuant to the mandatory retirement policy. | Time: April 1, 2022" query: "Pantoja's removal from elected office follows the approval of the new policy by the defendants." answer: yes