An Expert-Validated LLM Framework for Transforming Legal Procurement Texts into Actionable Data

Ivan SPADA [a,1], Roberto NAI [a], Davide AUDRITO [b],
Vittoria Sofia Margherita TRIFILETTI [c] and Emilio SULIS [a]
[a] *University of Turin - Department of Computer Science*
[b] *University of Bologna - Department of Legal Studies*
[c] *University of Turin - Department of Law*
ORCiD ID: Ivan Spada https://orcid.org/0009-0002-0459-1189, Roberto Nai
https://orcid.org/0000-0003-4031-5376, Davide Audrito
https://orcid.org/0000-0002-9239-5358, Vittoria Sofia Margherita Trifiletti
https://orcid.org/0009-0000-3155-758X, Emilio Sulis
https://orcid.org/0000-0003-1746-3733

Abstract. Legal and administrative sources typically describe procedural steps that are not recorded in structured data, which makes the application of process-oriented analysis particularly challenging. In this work, we present an approach that uses Large Language Models (LLMs) to extract events and dates from unstructured legal texts. The methodology is applied to a dataset of Italian procurement notices published since 2022 on the official EU platform, Tenders Electronic Daily (TED), demonstrating how LLMs can extract valuable information, such as administrative decisions adopted prior to tender publication. These elements are incorporated into existing event logs, thereby enhancing the quality of process analysis. A sample of the extracted data has been manually reviewed by legal experts to assess the relevance and correctness of the automated detection. The results suggest that this approach can help identify procedural steps hidden in free text, thereby supporting more complete and accurate representations of legal workflows.

Keywords. Legal Process Analysis, Event Log Enrichment, Textual Analysis in Procurement

1. Introduction

The growing availability of digital data has increased attention to legal research [3]. Legal and administrative procedures, such as public procurement, are documented through legislation, notices, and rulings, yet critical process information remains in unstructured text. While information systems record standardised events, relevant procedural details often appear only in free-text sections, hindering automated analysis.

This research applies Natural Language Processing (NLP) and Large Language Models (LLMs) [25] to extract events and dates from procurement notices on the EU

[1] Corresponding Author: Ivan Spada, ivan.spada@unito.it.

platform *Tenders Electronic Daily* (TED), enriching event logs for more complete process analyses. We propose a two-stage, expert-validated framework that combines LLM-based extraction with manual legal annotation to identify procedural steps absent from unstructured data.

Our contributions are: (i) a pipeline leveraging LLMs for extracting information from procurement notices with expert supervision through dual prompting and annotation; (ii) an expert-based assessment of LLM potential and limitations in legal contexts, particularly for boolean detection tasks. Section 2 reviews related work; Section 3 presents methodology; Section 4 describes the dataset; Section 5 presents results; Section 6 concludes describing the future work.

2. Related Work

The automated analysis of legal procedures can be approached through business process management (BPM) [12] and, in particular, process mining (PM) [36], which reconstructs workflows from event logs to identify bottlenecks [29,30], assess compliance [31], and optimise resources [27]. However, structured event data are rarely available in legal contexts [14], where procedural information often resides in textual sources such as regulations, rulings, or administrative documents [22]. Early computational models aimed to capture the normative structure of legal processes [15], while more recent hybrid approaches combine structured data and text to improve completeness [8,26].

Information extraction has long supported the creation of legal knowledge bases through rule-based and machine learning techniques [18]. Various NLP methods have been applied to Named Entity Recognition in legal documents [2], from early statistical models [11] to domain-specific corpora for Indian [17], English [4], and German [20] legal texts. Yet, these methods remain sensitive to domain variability and linguistic ambiguity.

The advent of LLMs has brought advances in summarisation, question answering [28], and event extraction [19]. In the legal domain, LLMs have been explored for judgment prediction [37], contract analysis [32], legal summarisation [10], harmonisation [5,6], and entity–relation extraction [21]. Despite these successes, concerns about bias and explainability highlight the need for expert supervision [9].

Building on our earlier work [35,27], which introduced a process-oriented framework for event extraction from tender notices, this study extends that approach with a two-stage expert-based validation strategy, achieving higher accuracy on a larger dataset.

3. Methodological Framework

This section outlines the methodological framework for identifying relevant events within legal documents. Our approach integrates Large Language Models (LLMs) with expert validation to extract structured data from unstructured legal texts. The framework consists of two main stages — extraction and evaluation — supported by iterative expert feedback, as shown in Figure 1.

Overview. The two-stage methodology (Figure 1) combines prompt-based event extraction with legal expert evaluation. Each stage includes an extraction phase — from prompt engineering to machine-readable linking — followed by manual expert assessment of accuracy and completeness. Two legal experts participated in the evaluation to ensure compliance with domain standards.

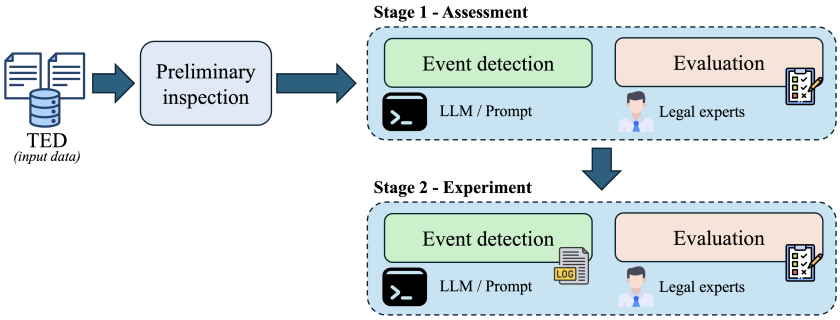

Figure 1. Overview of the two-stage expert-based methodology for event extraction and evaluation

Preliminary Inspection. An exploratory phase was conducted with domain experts, building on previous findings [35]. This phase refined the requirements and prompts by analysing a sample of legal documents and updating the extraction guidelines. Keyword and abbreviation filtering improved document selection (e.g., DL, DM, LR, DGR), ensuring that only notices likely to contain relevant event–date pairs were included. Collaboration between legal experts and computer scientists defined both domain-specific requirements and evaluation criteria, ensuring methodological consistency.

Stage 1. The assessment stage compared multiple prompting strategies across different LLMs to assess: (i) zero-shot, (ii) few-shot, and (iii) chain-of-thought (CoT) approaches. Manual evaluation of a subset of the dataset identified the best-performing model–prompt combination according to expert guidelines, enabling selection of the optimal configuration for large-scale testing.

Stage 2. In the experiment stage, the selected configuration was applied to a representative sample of TED documents. Multiple experts independently annotated the results, and inter-annotator agreement was measured using Cohen's κ [7]. This validation confirmed the reliability and domain validity of the extracted data.

4. From Methodology to Practice: The Case Study

This section presents the case study and its application of the proposed framework. The study analyses 27,841 Italian public procurement notices published on TED in 2022 [13]. The year 2022 was chosen because it followed the pandemic period and allowed completed procedures to be observed. While some sections of the notices follow a structured template with standardised fields, others contain free-text descriptions of procedural acts. In particular, the sections titled "complementary information" or "additional information" often include decrees and decisions marking the start of procurement procedures, which are the focus of this study.

Preliminary inspection. Following our framework, we conducted an exploratory analysis of TED documents with legal experts, extending our previous work [35] that used GPT-4o [33] for event–date extraction. Legal terminology and abbreviations from a regulatory database [34] (e.g., *DL*, *DM*, *LR*, *DGR*) were integrated to refine prompts and improve document selection. The filtering pipeline progressively reduced the dataset to 2,839 relevant notices by selecting documents with "complementary" or "additional information" sections, well-formed Italian dates, and at least one legal keyword or abbreviation. This ensured both legal relevance and computational efficiency.

Stage 1. Building upon [35], we extended the analysis from 2019 back to 2016 [16]. Three models were compared: GPT-4o [33], Llama-3.1-8B [24], and Claude-Sonnet-4 [1], each tested with zero-shot, few-shot, and chain-of-thought prompting. Manual evaluation on ten sample documents identified Claude-Sonnet-4 with CoT prompting as the most accurate and legally consistent configuration, particularly for handling temporal dependencies and domain-specific terminology.

Stage 2. The best-performing model–prompt combination was applied to a larger sample. Two legal experts independently annotated 100 instances, achieving high inter-annotator reliability (Cohen's κ), confirming the robustness and reproducibility of the extraction process.

Implementation details. The framework was implemented in Python 3.12 (64-bit) using `NLTK` for text preprocessing and APIs from OpenAI, Meta, and Anthropic. All materials—including code, guidelines, prompt strategies and datasets—are publicly available[2].

5. Results

Validation of the first extraction results. The first extraction was evaluated by two independent annotators using a three-label scheme (Yes, No, Doubts). The overall agreement reached a Cohen's κ of 0.576, which corresponds to a *moderate* level of inter-annotator reliability. This outcome suggests that, while the LLM-generated results are generally consistent, the presence of the Doubts category introduced some ambiguity into the annotation process, reflecting borderline or less clear-cut cases. After discussing and resolving the doubtful cases, the two legal experts proposed a final assessment of the events and dates correctly extracted by the LLM. Specifically, out of 121 cases, 89 answers were correct (74%), while the model was incorrect in 32 cases (24%). This outcome is unsatisfactory, thereby justifying a second annotation with a new prompt and a subset of notices to increase the percentage of successful cases.

Improved annotation guidelines. The two-stage annotation process enhanced guideline accuracy through iterative refinement and identification of previously undetected edge cases. Legal experts revised the annotation scheme from tripartite ("Yes/No/Doubt") to binary ("Yes/No") classification. This simplification eliminated the "doubt" category, which introduced unnecessary uncertainty without discriminative value. Establishing stricter, more explicit decision criteria from the outset reduced annotator uncertainty and improved inter-annotator agreement. Additionally, the target term set was expanded

[2] GitHub repository: https://github.com/roberto-nai/JURIX-2025

to include "provision" (in Italian, "disposizione"), thereby marking events of particular relevance that had previously been excluded.

Expert guidance enabled prompt refinement by selectively filtering concept qualifiers rather than entire categories. For instance, only legislative and law decrees were excluded, as they refer to general procurement regulations rather than individual notice lifecycles, whereas other decree types remained relevant to the analysis.

Assessment of strategies. The system prompt incorporates legal expert analysis of preliminary TED instances, specifying the extraction task, relevant keywords with abbreviations, and elements to exclude during event detection. This design translates legal reasoning into a computationally tractable format.

Three prompting strategies were evaluated. The zero-shot prompt (Prompt 1) provides a baseline that incorporates four expert-defined requirements, requesting JSON output with the "event" and "date" fields. The few-shot prompt (Prompt 2) extends this by adding three practical examples. The CoT prompt (Prompt 3) introduces explicit step-by-step reasoning for entity extraction in the process mining context, including a demonstration.

Manual evaluation on 10 randomly selected instances compared the three strategies across models with varying characteristics: *Claude-Sonnet-4* (Prompt 1: 8; Prompt 2: 7; Prompt 3: **9**), *GPT-4o* (Prompt 1: 7; Prompt 2: 7; Prompt 3: 8), *Llama-3.1-8B* (Prompt 1: 2; Prompt 2: 6; Prompt 3: 4). The optimal configuration was Claude-Sonnet-4 with CoT.

Validation of second extraction results. Inter-annotator agreement between two legal experts on 100 cases yielded Cohen's $\kappa = 0.896$, indicating *almost perfect* agreement. The model correctly identified 108 of 115 cases (94%), with 7 errors (6%), demonstrating substantial improvement over the initial extraction and confirming the effectiveness of the expert-based framework in enhancing accuracy and reducing hallucinations. These results represent a significant advancement for legal process digitalisation and optimisation.

6. Conclusion and Future Work

This work presents a two-stage framework for extracting legally meaningful events and dates from administrative texts. Applied to large-scale procurement data, it enables the identification of information useful for assessing public administrations' performance and provides an empirical basis for transparency and accountability analyses. From a legal perspective, the method also helps address systemic issues such as limited transparency and corruption. During evaluation, the LLM showed a partial ability to distinguish relevant from irrelevant information—correctly ignoring non-pertinent decrees or resolutions in several cases—while still struggling with more subtle distinctions, such as those between resolutions initiating or annulling a procedure.

A key insight is that legal meaning depends on institutional context and procedural relevance rather than on isolated terms. By incorporating expert validation, the approach demonstrates that legal knowledge is essential for guiding computational analysis and producing data-driven evaluations of procurement processes.

Future work will explore the semantic modelling of procurement notices using existing ontologies and their automatic generation in queryable, explanatory [23] formats using LLMs.

References

[1] Anthropic. Claude-Sonnet-4. `https://www.anthropic.com/claude/sonnet`, 2025. Accessed: 2025-09-10.

[2] Farid Ariai, Joel Mackenzie, and Gianluca Demartini. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *arXiv preprint arXiv:2410.21306*, 2024.

[3] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[4] Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. E-NER — an annotated named entity recognition corpus of legal text. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro, editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[5] Davide Audrito, Luigi Di Caro, Laura Genga, Rachele Mignone, Roberto Nai, Ivan Spada, Emilio Sulis, and Vittoria Margherita Sofia Trifiletti. Standardization and harmonization of law through automated process analysis and similarity techniques. In Laura Genga, Hugo A. López, and Emilio Sulis, editors, *Proceedings of the 1st International Workshop on Processes, Laws and Compliance co-located with 6th International Conference on Process Mining (ICPM 2024), Lyngby, Denmark, October 14, 2024*, volume 3850 of *CEUR Workshop Proceedings*, pages 34–45. CEUR-WS.org, 2024.

[6] Davide Audrito, Ivan Spada, Rachele Mignone, Emilio Sulis, and Luigi Di Caro. Towards semi-automating european legislative harmonisation analysis: A harmonised glossary for llm-based legal concept detection. *Computer Law & Security Review*, 58:106171, 2025.

[7] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[8] Matilde Contestabile, Chiara Ferrara, Alberto Giovannetti, Giovanni Parrillo, and Andrea Vandin. The prolific dataset: Leveraging llms to unveil the italian lawmaking process, 2025.

[9] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024.

[10] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, pages 1–44, 2024.

[11] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. Named entity recognition and resolution in legal text. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, pages 27–43. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[12] Marlon Dumas, Marcello La Rosa, Jan Mendling, and Hajo A. Reijers. *Fundamentals of Business Process Management*. Springer, 2018.

[13] European Union. Tenders electronic daily (ted) (csv subset) – public procurement notices. `https://data.europa.eu/data/datasets/ted-csv?locale=en`, 2024.

[14] Dirk Fahland. Extracting and pre-processing event logs. *arXiv preprint arXiv:2211.04338*, 2022.

[15] Governatori, Guido et al. Logic and the law: philosophical foundations, deontics, and defeasible reasoning. *Handbook of Deontic Logic and Normative Reasoning*, 2:655–760, 2021.

[16] Italian Government. Legislative decree no. 50/2016 and legislative decree no. 36/2023 on the public procurement code, 2023. The procurement code introduced by Legislative Decree No. 50/2016 remained in force until its replacement in March 2023 (Decree No. 36/2023).

[17] Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. Named entity recognition in Indian court judgments. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro, editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[18] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. Recent advances in named entity recognition: A comprehensive survey and comparative study, 2024.

[19] Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artif. Intell. Rev.*, 57(9):260, 2024.

[20] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. A dataset of German legal documents for named entity recognition. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri,

Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France, May 2020. European Language Resources Association.

[21] Shiye Li and Li Yi. A few-shot entity relation extraction method in the legal domain based on large language models. In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, DEAI '24, page 580–586, New York, NY, USA, 2024. Association for Computing Machinery.

[22] Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. Better call gpt, comparing large language models against lawyers. *arXiv preprint arXiv:2401.16212*, 2024.

[23] Rosa Meo, Roberto Nai, and Emilio Sulis. Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next? In Silvia Chiusano, Tania Cerquitelli, and Robert Wrembel, editors, *Advances in Databases and Information Systems - 26th European Conference, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings*, volume 13389 of *Lecture Notes in Computer Science*, pages 25–34. Springer, 2022.

[24] Meta. Llama-3.1-8B. `https://huggingface.co/meta-llama/Llama-3.1-8B`, 2024. Accessed: 2025-09-10.

[25] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[26] Roberto Nai, Ishrat Fatima, Gabriele Morina, Emilio Sulis, Laura Genga, Rosa Meo, and Paolo Pasteris. AI applied to the analysis of the contracts of the italian public administrations. In Fabrizio Falchi, Fosca Giannotti, Anna Monreale, Chiara Boldrini, Salvatore Rinzivillo, and Sara Colantonio, editors, *Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023), Pisa, Italy, May 29-30, 2023*, volume 3486 of *CEUR Workshop Proceedings*, pages 255–260. CEUR-WS.org, 2023.

[27] Roberto Nai, Emilio Sulis, Davide Audrito, Vittoria Margherita Sofia Trifiletti, Rosa Meo, and Laura Genga. Leveraging process mining and event log enrichment in european public procurement analysis: a case study. *Computer Law & Security Review*, 57:106144, 2025.

[28] Roberto Nai, Emilio Sulis, Ishrat Fatima, and Rosa Meo. Large language models and recommendation systems: A proof-of-concept study on public procurements. In Amon Rapp, Luigi Di Caro, Farid Meziane, and Vijayan Sugumaran, editors, *NLDB 2024, Turin, Italy, June 25-27, 2024, Proceedings*, volume 14763 of *Lecture Notes in Computer Science*, pages 280–290. Springer, 2024.

[29] Roberto Nai, Emilio Sulis, and Laura Genga. Automated analysis with event log enrichment of the european public procurement processes. In Tiago Prince Sales, João Araújo, José Borbinha, and Giancarlo Guizzardi, editors, *Advances in Conceptual Modeling - ER 2023 Workshops, JUSMOD, Lisbon, Portugal, November 6-9, 2023, Proceedings*, volume 14319 of *Lecture Notes in Computer Science*, pages 178–188. Springer, 2023.

[30] Roberto Nai, Emilio Sulis, Rosa Meo, Francesco Gorgerino, Gabriella Margherita Racca, and Laura Genga. Process mining on a public procurement dataset: A case study. In Rosa Meo and Fabrizio Silvestri, editors, *International Workshops of ECML PKDD 2023, Turin, Italy, Sept. 18-22, 2023, Revised Selected Papers, Part I*, volume 2133 of *Communications in Computer and Information Science*, pages 477–492. Springer, 2023.

[31] Roberto Nai, Emilio Sulis, Paolo Pasteris, Mirko Giunta, and Rosa Meo. Exploitation and merge of information sources for public procurement improvement. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part I*, 2022.

[32] Savinay Narendra, Kaushal Shetty, and Adwait Ratnaparkhi. Enhancing contract negotiations with LLM-based legal document comparison. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis, editors, *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 143–153, Miami, FL, USA, November 2024. Association for Computational Linguistics.

[33] OpenAI. GPT-4o. `https://openai.com/research/gpt-4o`, 2024. Accessed: 2025-09-10.

[34] PIM - Portale della Regione Lombardia. Legend of the abbreviations used in the regulatory database. `https://www.pim.mi.it/legenda-dei-termini-utilizzati-nella-sezione-normativa`, 2025. Accessed: 2025-10-20.

[35] Ivan Spada, Emilio Sulis, Davide Audrito, Vittoria Margherita Sofia Trifiletti, and Roberto Nai. Aug-

menting public procurement event logs with large language models: a legal process mining approach. In *Proceedings of the Joint Ontology Workshops (JOWO) - Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025)*, 2025.

[36] Wil M. P. van der Aalst. *Process Mining: Data Science in Action*. Springer, 2nd edition, 2016.

[37] Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Kumar Nigam, Shouvik Kumar Guha, Koustav Rudra, and Kripabandhu Ghosh. LLMs - the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on indian court cases. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12451–12474. Association for Computational Linguistics, 2023.