

Augmenting Public Procurement Event Logs with Large Language Models: a Legal Process Mining Approach

Ivan Spada^{1,*†}, Emilio Sulis^{1†}, Davide Audrito^{2†}, Vittoria Margherita Sofia Trifiletti^{3†} and Roberto Nai^{1,*†}

¹Computer Science Department, University of Turin, Corso Svizzera 185, 10149, Torino (IT)

²Department of Legal Studies, University of Bologna, Via Zamboni 27, 40126, Bologna (IT)

³Department of Law, University of Turin, C.so Lungo Dora Siena 100, 10153, Torino (IT)

Abstract

In process mining, the construction of an event log is the foundation for all subsequent analysis. The enrichment of event logs with information extracted from unstructured data is a promising area with great potential that remains largely unexplored. Nevertheless, recent advances in Large Language Models present new opportunities in the field. This paper proposes a framework to extract valuable information from large volumes of textual data. In particular, we demonstrate our approach through a proof-of-concept in the legal domain, analysing textual information from tender notices. We demonstrate how LLMs help identify events and temporal data, enriching event logs and enabling more comprehensive process analysis. Two domain experts in the legal field provide their expertise, e.g. validating the approach, as well as improving the text processing task. Finally, the results obtained are discussed to explore the integration of LLMs to support process mining through the identification of “hidden” information from unstructured data, as in the case of legal texts.

Keywords

Process Mining, Large Language Models, Event log enrichment, Legal text

1. Introduction

Process mining has emerged as a foundational methodology for analysing and optimising business processes by systematically examining event logs extracted from information systems. Within the business process lifecycle, the construction of comprehensive event logs constitutes the critical first step upon which all subsequent analysis is based. The integrity of process mining outcomes, including insights into process performance, compliance verification, and efficiency assessments, is fundamentally contingent on the accuracy, completeness, and granularity of these event logs [1].

Conventional approaches primarily derive event logs from structured data repositories, such as Enterprise Resource Planning systems. However, this established paradigm neglects the substantial volume of potentially valuable process information embedded within unstructured textual data. Recent advancements in Natural Language Processing (NLP), particularly the emergence of Large Language Models (LLMs), have created new opportunities for extracting meaningful insights from unstructured textual data. These sophisticated models, trained on vast and diverse corpora, demonstrate remarkable capabilities in context handling, pattern recognition, and entity identification within complex documents.

This paper presents a systematic investigation into the integration of LLMs within process mining workflows. Our research focuses specifically on enhancing event log fidelity and comprehensiveness through the extraction of information embedded in unstructured textual sources. In particular, we propose a framework to investigate the automated extraction of new events and dates from legal

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), September 8-9, 2025, Catania, Italy

*Corresponding author.

†These authors contributed equally.

✉ ivan.spada@unito.it (I. Spada); emilio.sulis@unito.it (E. Sulis); davide.audrito2@unibo.it (D. Audrito); vittoriarmargheritasofia.trifiletti@unito.it (V. M. S. Trifiletti); roberto.nai@unito.it (R. Nai)

ORCID 0009-0002-0459-1189 (I. Spada); 0000-0003-1746-3733 (E. Sulis); 0000-0002-9239-5358 (D. Audrito); 0009-0000-3155-758X (V. M. S. Trifiletti); 0000-0003-4031-5376 (R. Nai)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

text. Such information can be incorporated into event logs, enhancing the accuracy of the following process analysis. The proposed framework highlights the potential of combining LLM-driven extraction techniques with traditional process mining methodologies, paving the way for more comprehensive and data-informed process optimisation strategies. In a practical case study, we focus on a specific application within the legal domain. Specifically, we analyse tender notices that contain rich, yet unstructured, information about public procurement processes. As a proof-of-concept (POC), we focus on the complementary information contained in Tenders Electronic Daily (TED) documents from 2022. The research question investigates how LLMs can extract meaningful process-related information from unstructured legal texts and enhance the completeness of event logs in public procurement process analysis.

The paper is organised as follows: Section 2 synthesises pertinent literature regarding the utilisation of unstructured data in process mining methodologies. Section 3 delineates our proposed framework, employed in a case study described in Section 4. Section 5 elucidates the experimental configuration and preliminary findings, followed by a comprehensive analytical discussion in Section 6. Finally, Section 7 presents concluding observations and implications for future research trajectories.

2. Related work

Process Mining (PM) is a research area that focuses on extracting knowledge from event logs to analyse and improve business processes [2]. The discipline combines techniques from data mining, business process management, and machine learning to discover, monitor, and optimise workflows based on real execution data. PM techniques include methods for event log generation, process discovery, conformance checking, and predictive analytics [3].

Despite traditional PM methods relying on structured event logs, the increasing availability of unstructured data sources led to a growing interest in combining NLP techniques and process analysis [4]. Therefore, NLP techniques can be used in PM in case of textual information related to business processes, i.e. documents which may contain valuable process-related knowledge that can be extracted and structured into event logs. Such techniques can also be applied to legal processes, as confirmed by the growing attention on process mining in the legal field [5].

The adoption of LLMs, such as BERT, GPT, and transformers, has recently significantly focused on process modelling [6, 7], abstraction techniques [8], semantic-aware process mining tasks [9], event abstraction [10]. A recent work investigates the adoption of LLM to create SQL queries for the extraction of event logs from the database [11]. In particular, the identification of events within textual data can be performed with several methods using rule-based linguistic techniques (e.g., a recent work investigates event log enrichment from unstructured legal text with regular expressions [12]). In this work, we propose the adoption of LLMs to perform the task.

A crucial aspect of NLP-driven PM is the enrichment of event logs by integrating information extracted from unstructured text. Beyond traditional event identification, event log enrichment involves associating extracted entities with structured process representations, enhancing timestamps, detecting missing events, and linking unstructured descriptions to predefined process activities. We address this challenge with LLMs-based automatic extraction of events and dates from text, within a human-in-the-loop validation framework [13] based on legal domain experts involved in the research.

The increasing adoption of these techniques across various domains further demonstrates the potential of NLP-enhanced process mining to address complex, real-world challenges. Process mining techniques have been increasingly applied across diverse domains, including education [14, 15], administrative and legal settings [16, 17], public procurement [18], fraud detection [19], and regulatory compliance [20], often in combination with NLP methods [21, 22]. These cross-domain applications demonstrate the flexibility of PM methodologies and highlight the growing relevance of integrating textual information into process-aware analyses. Building on this trend, recent efforts have begun exploring the application of NLP-enhanced PM in healthcare scenarios, where the complexity and critical nature of clinical workflows require explainability [23], traceability, and robust integration of structured and unstructured

data.

3. Methodological framework

In this section, we introduce a methodological framework for enriching process mining event logs through the extraction of events from unstructured legal documents, which has been implemented in a POC.

The proposed framework consists of a pipeline designed to transform unstructured textual information into structured event log entries compatible with standard process mining algorithms. The idea we started with is to extract a new event with a date, but here we propose key steps that can also be applied to other types of textual information to be extracted. This pipeline encompasses document preprocessing, information extraction, validation mechanisms, and description of newly identified process elements. We detail distinct modules of the methodological pipeline, as illustrated in Figure 1.

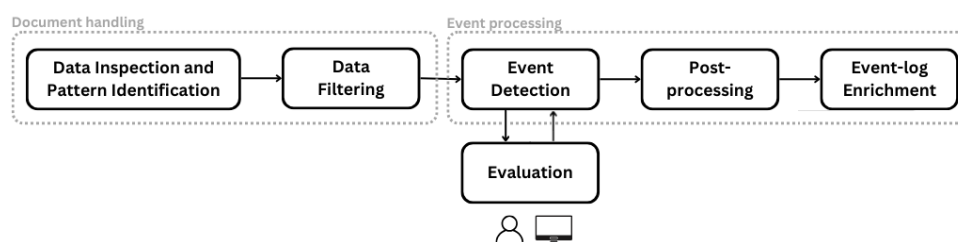


Figure 1: A general framework for automated extraction of information from unstructured data for the enrichment of event log for process mining

3.0.1. Data Inspection and Pattern Identification.

Following the identification of the domain and initial document corpus, a preliminary qualitative data inspection conducted by domain experts facilitates the identification of specific cases, requirements, and patterns. Through this analytical process, relevant textual segments containing pertinent information emerge, along with additional requirements such as the presence of keywords that identify specific events or activities beneficial for inclusion in the log file for subsequent process mining tasks.

3.0.2. Data Filtering.

Once the textual segments to process in the domain-specific requirements are identified, it may be necessary to apply filtering mechanisms for reducing dimensionality and consequently resource consumption. For instance, events may fall exclusively within a subset of possible activities that can be identified through natural language processing methodologies, including stemming and lemmatisation. With respect to dates, those outside the temporal range of interest may be excluded.

3.0.3. Event Detection.

This module encompasses the extraction of data of interest, e.g. the name of the event and the corresponding date pairs from the (filtered) dataset. When data exhibit standard structures, classical natural language processing methodologies may be employed (e.g., regular expressions). Alternatively, LLMs or comparable approaches can be used when the unstructured nature of the data necessitates more advanced semantic context management.

3.0.4. Evaluation on a sample of cases.

The results obtained must undergo an evaluation to assess the performance of the methodology and the selected requirements relative to the domain, dataset, and data type utilised. The evaluation may

proceed through automatic or manual approaches contingent upon the existence of a gold standard or alternative automatic evaluation methods; otherwise, manual inspection would confer enhanced experimental relevance. Suggestions collected at this stage allow directing the final extraction and provide useful suggestions for improving the data through post-processing.

3.0.5. Post-processing.

Results obtained in the LLM extraction phase may contain noise, necessitating further investigation or removal. For instance, the corresponding documents containing keywords and dates within a specific range might also encompass event-date pairs irrelevant to log file enrichment.

3.0.6. Event log Creation and Enrichment.

Upon acquisition of new information, as in the case of new event-date pairs, incorporation into either new log files or existing ones becomes necessary. In the latter scenario, newly inserted entries must maintain consistent formatting, contain compatible case identifiers (which could align with existing records in the log file belonging to the same case), and establish default values for potentially incomplete fields.

4. Proof-of-concept

A case study has been selected to apply the proposed framework. This section describes the data, the practical steps, and an example of utility in process mining. In particular, we implemented a proof-of-concept focused on public procurement processes. Specifically, we relied on a previous work in which we already extracted an event log from data with 6 events from the publication of the notice to the closing of the contract [12]. The goal of the actual POC is to extract new data, when present, that is related to the procedure under consideration.

4.1. Dataset

We utilised a corpus of TED documents from 2022, which represent official public procurement notices from the European Union¹. These documents were selected due to their standardised yet linguistically complex nature, containing rich procedural information often embedded within legal terminology and narrative text.

4.1.1. Data on tenders.

Each tender is identified by an alphanumeric code referred to as the *document-number*, which serves as the key value. The TED dataset includes several relevant attributes: the *sector* each tender falls under (Services, Works, or Supplies); the *NUTS* (Nomenclature of Territorial Units for Statistics) code of the contracting authority responsible for issuing the tender notice; the *type* of contracting authority (such as a Ministry, European Institution, or Regional/local authority); and the *amount* associated with the tender.

4.1.2. The legal process.

We specifically focused on regional authorities within the Italian cases. The legal process of public procurements includes the following five main events. The process starts with the event “publication” of the tender, followed by the “participation” phase, during which individual entities submit their bids. After this step, the evaluation process leads to the “award”, where the contract is assigned to the selected bidder. Subsequently, the contract becomes legally effective (“contract-start”), and the process concludes upon reaching the deadline specified in the tender notice (“contract-end”).

¹TED portal: <https://data.europa.eu/data/datasets/ted-csv?locale=en>

4.1.3. Overview of the dataset.

The full dataset explored in our proofs-of-concept concerns 27,841 tenders, available in PDF and XML format. As our main interest is to develop an automated method to extract new activities and the corresponding dates from the unstructured text of public tenders, we manage the TED dataset in XML format to extract data of interest. The documents are in the Italian language and contain sections such as “contracting authority”, “subject”, “legal, economic, financial, and technical information”, “procedure”, and “additional information”. A recent work provides a broader description of the dataset [16]. In our POC, we focus on a subset of the Italian tender dataset for the year 2022².

4.2. Methods and Technologies

This subsection describes the application of the framework proposed in Section 3 to the specific legal case of interest. The development of the modules in Figure 1 is detailed below according to the two macro sets of modules related to data management and event extraction for insertion into logfiles.

4.2.1. Document handling.

Upon inspection of the previously described dataset, the information useful for enriching existing logfiles is contained in the “additional information” section, which corresponds to the <INFO_ADD> tags aggregated within <COMPLEMENTARY_INFO> in XML formats. Considering the legend of acronyms and abbreviations used in the Regulations database section³, a set of stems⁴ and acronyms⁵ were defined for entries.

The documents containing complementary information do not always exist, as this is an optional section, number 10781. Among these, 3505 documents contain dates in the Italian format “dd/mm/yyyy”. Subsequently, the TED documents containing at least one of the selected stems in their complementary information were filtered, yielding 2352 instances on which to experiment. At this point, a mapping was performed in a separate document with the caseid, the TED document identifier, and the related complementary information.

4.2.2. Event processing and evaluation.

The unstructured nature of relevant information contained in TED documents and the flexibility with which events can be described in the extracted sections led to the choice of leveraging an LLM-based approach for the event detection task. We select for the experiment GPT-4o⁶, as it is considered one of the best-performing models.

The prompt (Appendix A) requests the extraction of event-date pairs from documents following six rules useful for describing the task without making the prompt too specific. Subsequently, the complementary information of the instance is inserted and the output requested is a table with the following columns: “event” and “date.” This method reduces the randomness of the generated output structure and produces a result that is both machine-readable and easy for humans to interpret. The results, in *markdown* format, were stored in a mapping with the respective caseids.

In total, 2959 event entries (called “name-of-activity”) and dates were obtained. To clean the outputs and obtain entries to be annotated with actual events, entries with “name-of-activity” not contained in the complementary information were removed, thus obtaining 1015 entries. According to legal experts,

²The 2022 Italian TED dataset used is available at <https://bit.ly/46JZe37>

³Legend of acronyms/abbreviations used in the Normative section database: “tipologia norma”. Available at: <https://www.pim.mi.it/legenda-dei-termini-utilizzati-nella-sezione-normativa>

⁴[“deliber”, “det”, “determin”, “dirett”, “provved”, “decret”, “atti”, “comunic”, “atto”, “ordin”]

⁵[“DL”, “DM”, “DCM”, “DPCM”, “DLGS”, “DAR”, “DCR”, “DGR”, “DUPCR”, “DCCR”, “DPGR”, “DDS”, “DSGR”, “DDUO”, “DDG”, “DCIPE”, “CU”, “OCDPC”, “DSM”, “DCMM”, “DCoM”, “DA”, “DCC”, “DCUC”, “DGC”, “DCS”, “DD”, “DRA”, “DDP”, “DPP”, “DCP”, “DGP”, “CdG”, “DComP”]

⁶OpenAI, GPT-4o. Available at: <https://openai.com/index/hello-gpt-4o>

entries with dates prior to January 1st, 2019, were ignored, as they cannot refer to the case of interest because they are too far prior to the advertisement of the notice.

Evaluation. To evaluate the effectiveness of the experiment, two legal experts explored 50 instances each, selected randomly. Since each instance corresponds to a caseid, each document may contain multiple event-date pairs within the complementary information, producing a total of 121 entries. Doubtful cases helped clarify aspects of extraction, as well as identify new possible events to be extracted.

The guidelines⁷ describe the annotation task, which requires, for each entry, indicating a label and any notes. The labels assume the value "correct" if the extracted pair is suitable for enriching the log file, "doubt" if there is a problem in the extraction with doubts to be identified in the notes, and "error" if the extraction is considered as incorrect. The considerations of doubtful cases and errors are then discussed to propose improvements in the extraction or post-processing phase.

5. Results

This section presents the main outcomes of the different phases of the POC. We focus initially on the first results of the extraction evaluation, provided by two domain experts, and then dwell on the extraction carried out by the LLM. Finally, we provide an example of how information extracted from language models can significantly enrich the log for legal process analysis.

5.1. Evaluation results

The evaluation of the sample of LLMs results offers interesting insights into several aspects. In fact, the LLM provided the correct result (71.9%). The "error" cases were about 10%, a not-so-negligible percentage. Nevertheless, a manual inspection of error cases demonstrates how almost all cases can be managed with post-processing steps of the LLMs results. For instance, the most common errors were using an inappropriate portion of text that did not contain the keywords used in the identification phase. Therefore, a script in post-processing has been sufficient to remove the occurrences that did not contain the keywords.

With "doubt" cases, we obtain meaningful insights. First, the manual inspection enables the identification of new keywords (e.g., "Disposizione"), which can be added to the set of keywords for a new extraction. Additionally, a verification reveals the identification of new events (e.g., "request for clarification"). We have opted not to include this event in the actual process analysis, but it could be considered for future work. Moreover, it was observed that in a few cases, the LLM identifies two different dates for the same event. This is not an error, as the legal text contains two similar events occurring on distinct but closely related dates. However, the heuristic considers the earlier date as the most accurate.

Finally, we consider it physiological that there may be some errors during automatic extraction, but the proposed arrangements (initial keywords to identify the paragraph, stemming, and post-processing) allow us to achieve a very positive result by minimising any errors. In addition, the domain experts noticed from the LLM results new events of interest with the corresponding date (e.g., request for clarification), which, however, is not considered in the current POC as we are interested in the administrative decision preceding the publicization of the tender.

5.2. Event log enrichment with LLMs

According on the annotators' feedback, entries were removed if their "name-of-activity" field did not contain one of the selected stems (after legal acts had been excluded from consideration). As already mentioned, dates prior to 2019 were disregarded as they were not relevant to this study .

⁷The annotation guidelines and the annotated spreadsheet are available at <https://bit.ly/46JZe37>

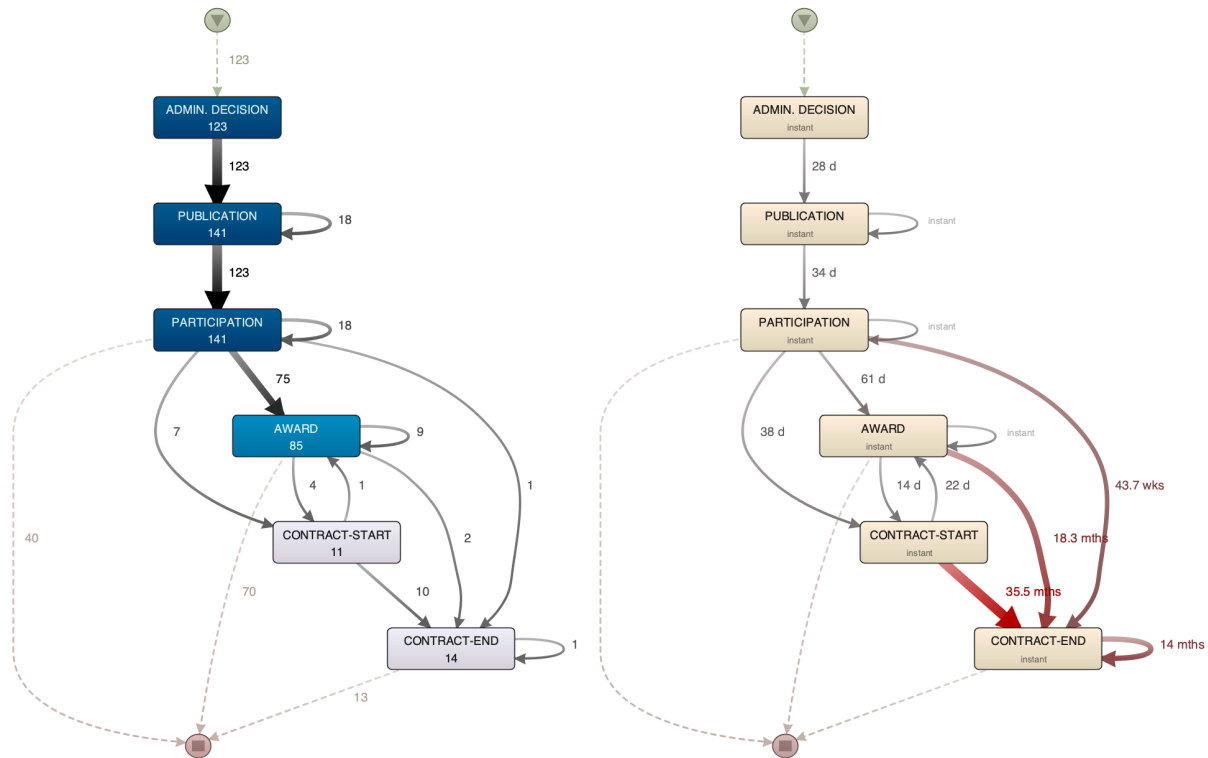


Figure 2: Diagrams of the legal process for activity duration (a) and performance (b) with the new event obtained with LLMs (Administrative decision).

In cases where multiple dates were detected for the same caseid, the annotators suggested inserting only the oldest date in the log file, as it is the only one of interest in the process. In this POC, we add the new event "Administrative decision" to the event log, which can be explored through traditional process discovery algorithms.

5.3. Process Discovery

To conclude the presentation of our POC, we propose the legal process discovered with the log enrichment. The extracted event concerns a subset of cases, as not all TED texts contain such a section, and not all sections contain the event of interest. By filtering the log for cases starting with the new event, we obtain the new activity diagrams that include 6 events, both for activity frequency and performance (the median duration between the activities weights arcs) as depicted in Figure 2.

6. Discussion

In conducting the present analysis, the full range of prodromal acts that occur before the tender process was identified, with it being noted that the legal document initiating this process may be designated in various ways depending on the procurement type and the issuing authority. This variability is attributable to the fact that different sectors and public bodies operate under distinct Procedural Requirements and Legal Traditions. As a consequence, the fundamental act - the formal decision to commence the tender process - can be labelled in different ways. Examples of such designations include "Authorisation to Proceed", "Decision to Contract", "Procurement Decision", "Awarding Decision", and "Notice of Intended Procurement", reflecting the nuanced influence of legal and administrative practices on the naming of these essential documents.

The utilisation of large language models for the extraction of data from legal texts constitutes a substantial progression in optimising the efficacy of legal document analysis. The evaluation task under scrutiny highlights the model's capacity for rapid extraction of key information, including events and

dates. However, the presence of errors in the outputs of LLMs necessitates a cautious approach, given the inherent complexity and nuanced nature of legal language, which can lead to misinterpretations. While the risk of embedded biases within the model's training data appears minimal in this context, it remains essential that automated data extraction undergoes comprehensive human review to ensure accuracy and to mitigate any potential for erroneous or discriminatory outcomes. Furthermore, the capability to accurately ascertain the point in time at which preliminary actions were commenced during the tender process is of immense legal significance. This enables a more profound comprehension of the procedural issues related to adjudication and facilitates the identification of any inconsistencies on the part of the public administration in adhering to the legally mandated time frame between the prodromal steps and adjudication. An interesting additional consideration concerns the evaluation phase and the role of domain experts involved in the proposed framework. As the slight disagreement between the two annotators in the evaluation stage on doubt cases showed, the legal domain presents considerable complexity. The human-in-the-loop approach therefore, still proves to be the correct one for this type of analysis, for which full automation does not seem easily achievable at present.

7. Conclusions and future work

We introduced a framework that enhances process mining by leveraging LLMs to extract meaningful information from unstructured data. Through a case study in the legal domain, we demonstrated how language models can identify events and data in tender notices, enriching event logs for enhancing process analysis. Our results highlight the potential of integrating LLMs exploitation into automated process analysis to improve the accuracy and depth of event representations. In the future, we intend to apply this technique to conduct a comprehensive analysis of the distinctions among various public procurement procedures, taking into account the new implications introduced by the updated Italian procurement code. First, we aim to conduct a more extensive annotation task involving a larger number of legal cases. In addition, we intend to systematically analyse instances of annotator disagreement and compute inter-annotator agreement metrics to assess the reliability and consistency of the annotations, based on the guidelines proposed and tested in the current work. Moreover, we plan to compare GPT-4o with other competitor models among both open- and closed-source options. Finally, it is of interest to extract new events from the data, as we noted in the evaluation phase (request for clarification) and consider legal documents in different languages.

Acknowledgements

The research work in this article was partially conducted as part of the following projects: the Circular Health European Digital Innovation Hub (CHEDIH) - Grant Agreement n. 101083745; PiemontAIs - PR FESR 2021/2027 Grant Agreement n. 187173.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to perform a grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] M. Dumas, M. L. Rosa, J. Mendling, H. A. Reijers, *Fundamentals of Business Process Management*, Springer, 2018. doi:10.1007/978-3-662-56509-4.
- [2] W. M. P. van der Aalst, *Process Mining - Data Science in Action*, Springer, 2016. doi:10.1007/978-3-662-49851-4.

- [3] W. M. P. van der Aalst, J. Carmona (Eds.), *Process Mining Handbook*, volume 448 of *LNBIP*, Springer, 2022. doi:10.1007/978-3-031-08848-3.
- [4] B. Estrada-Torres, A. del-Río-Ortega, M. Resinas, Mapping the landscape: Exploring large language model applications in business process management, in: H. van der Aa, D. Bork, R. Schmidt, A. Sturm (Eds.), *Enterprise, Business-Process and Information Systems Modeling*, volume 511 of *LNBIP*, Springer, 2024, pp. 22–31. doi:10.1007/978-3-031-61007-3_3.
- [5] L. Genga, H. A. López, E. Sulis, Emerging challenges in legal informatics from machine learning to llms-preface to the proceedings of the 1st plc workshop, in: 1st International Workshop on Processes, Laws and Compliance, CEUR-WS, 2024. URL: <https://ceur-ws.org/Vol-3850/preface.pdf>.
- [6] H. Kourani, A. Berti, D. Schuster, W. M. P. van der Aalst, Process modeling with large language models, in: H. van der Aa, D. Bork, R. Schmidt, A. Sturm (Eds.), *Enterprise, Business-Process and Information Systems Modeling*, Springer Nature Switzerland, Cham, 2024, pp. 229–244.
- [7] A. Norouzifar, H. Kourani, M. Dees, W. M. P. van der Aalst, Bridging domain knowledge and process discovery using large language models, in: K. Gdowska, M. T. Gómez-López, J.-R. Rehse (Eds.), *BPM Workshops*, Springer Nature Switzerland, Cham, 2025, pp. 44–56.
- [8] A. Berti, D. Schuster, W. M. P. van der Aalst, Abstractions, scenarios, and prompt definitions for process mining with llms: A case study, in: J. D. Weerd, L. Pufahl (Eds.), *Business Process Management Workshops - BPM 2023 International Workshops*, Utrecht, The Netherlands, September 11-15, 2023, Revised Selected Papers, volume 492 of *LNBIP*, Springer, 2023, pp. 427–439. doi:10.1007/978-3-031-50974-2_32.
- [9] A. Rebmann, F. D. Schmidt, G. Glavaš, H. van Der Aa, Evaluating the ability of llms to solve semantics-aware process mining tasks, in: 2024 6th International Conference on Process Mining (ICPM), 2024, pp. 9–16. doi:10.1109/ICPM63005.2024.10680677.
- [10] E. Brzychczy, K. Kluza, L. Szala, Enhancement of low-level event abstraction with large language models (llms), in: K. Gdowska, M. T. Gómez-López, J. Rehse (Eds.), *BPM Workshops*, Krakow, Poland, Sept. 1-6, 2024, volume 534 of *Lecture Notes in Business Information Processing*, Springer, 2024, pp. 209–220. doi:10.1007/978-3-031-78666-2_16.
- [11] V. S. Dani, M. Dees, H. Leopold, K. Busch, I. Beerepoot, J. M. E. M. van der Werf, H. A. Reijers, Event log extraction for process mining using large language models, in: M. Comuzzi, D. Grigori, M. Sellami, Z. Zhou (Eds.), *Cooperative Information Systems - 30th International Conference, CoopIS 2024*, Porto, Portugal, November 19-21, 2024, Proceedings, volume 15506 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 56–72. doi:10.1007/978-3-031-81375-7_4.
- [12] R. Nai, E. Sulis, L. Genga, Automated analysis with event log enrichment of the european public procurement processes, in: T. P. Sales, J. Araújo, J. Borbinha, G. Guizzardi (Eds.), *Advances in Conceptual Modeling - ER 2023 Workshops, JUSMOD*, Lisbon, Portugal, November 6-9, 2023, Proceedings, volume 14319 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 178–188. doi:10.1007/978-3-031-47112-4_17.
- [13] C. Fernandez-Llatas, *Interactive Process Mining in Healthcare: An Introduction*, Springer International Publishing, Cham, 2021, pp. 1–9. doi:10.1007/978-3-030-53993-1_1.
- [14] R. Nai, E. Sulis, L. Genga, Enhancing e-learning effectiveness: a process mining approach for short-term tutorials, *J. Intell. Inf. Syst.* 62 (2024) 1773–1794. doi:10.1007/s10844-024-00874-9.
- [15] R. Nai, E. Sulis, E. Marengo, M. Vinai, S. Capecchi, Process mining on students’ web learning traces: A case study with an ethnographic analysis, in: O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. A. Perifanou, T. Papatoma (Eds.), *Responsive and Sustainable Educational Futures - 18th EC-TEL 2023*, Aveiro, Portugal, September 4-8, 2023, Proceedings, volume 14200 of *LNCS*, Springer, 2023, pp. 599–604. doi:10.1007/978-3-031-42682-7_48.
- [16] R. Nai, E. Sulis, R. Meo, Ith: an open database on italian tenders 2016–2023, *Scientific Data* 11 (2024) 1452. doi:10.1038/s41597-024-04342-5.
- [17] R. Nai, E. Sulis, P. Pasteris, M. Giunta, R. 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. doi:10.1007/978-3-031-23618-1_6.

- [18] R. Nai, E. Sulis, R. Meo, F. Gorgerino, G. M. Racca, L. Genga, Process mining on a public procurement dataset: A case study, in: R. Meo, F. Silvestri (Eds.), 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*, Springer, 2023, pp. 477–492. doi:10.1007/978-3-031-74630-7_35.
- [19] R. Nai, E. Sulis, R. Meo, Public procurement fraud detection and artificial intelligence techniques: a literature review, in: D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. D. Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, N. Troquard (Eds.), Companion Proceedings of the 23rd EKAW Conference, Bozen-Bolzano, Italy, September 26-29, volume 3256 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: <http://ceur-ws.org/Vol-3256/km4law4.pdf>.
- [20] R. Nai, E. Sulis, D. Audrito, V. M. S. Trifiletti, R. Meo, L. Genga, Leveraging process mining and event log enrichment in european public procurement analysis: a case study, *Computer Law & Security Review* 57 (2025) 106144. doi:<https://doi.org/10.1016/j.clsr.2025.106144>.
- [21] R. Nai, R. Meo, G. Morina, P. Pasteris, Public tenders, complaints, machine learning and recommender systems: a case study in public administration, *Comput. Law Secur. Rev.* 51 (2023) 105887. doi:10.1016/J.CLSR.2023.105887.
- [22] R. Nai, E. Sulis, I. Fatima, R. Meo, Large language models and recommendation systems: A proof-of-concept study on public procurements, in: A. Rapp, L. D. Caro, F. Meziane, V. Sugumaran (Eds.), NLDB 2024, Turin, Italy, June 25-27, 2024, Proceedings, volume 14763 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 280–290. doi:10.1007/978-3-031-70242-6_27.
- [23] R. Meo, R. Nai, E. Sulis, Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next?, in: S. Chiusano, T. Cerquitelli, R. Wrembel (Eds.), ADBIS, Turin, Italy, Sept. 5-8, 2022, Proceedings, volume 13389 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 25–34. doi:10.1007/978-3-031-15740-0_3.

A. Appendix: Prompt

The full prompt, used for the final extraction following the outcome of the interaction with the domain experts.

Extract date-event pairs from text, following these rules:

1. Only include dates in dd/mm/yyyy format
2. Ignore incomplete dates (missing day, month, or year)
3. Event names contain the type of document and document number
4. Present results in a markdown table
5. Each row contains the event name and its associated date
6. No additional text before or after the table

Input text:

```
""
complementary_info
""
```

Table format:

```
| Event | Date |
|----|----|
| event name | dd/mm/yyyy |
```