From Clinical Event Logs to Narrative-Based Outcome Prediction with Large Language Models: A Case Study

Roberto Nai[1][0000−0003−4031−5376], Emilio Sulis[1][0000−0003−1746−3733], Laura Genga[2][0000−0001−8746−8826], and Adriana Boccuzzi[3][0000−0002−5789−1997]

[1] Computer Science Department - University of Turin
{roberto.nai,emilio.sulis}@unito.it
[2] School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands l.genga@tue.nl
[3] A.O.U. San Luigi Hospital, Regione Gonzole 10, 10043, Orbassano, Italy
adriana.boccuzzi@unito.it

Abstract. Managing patient flow in emergency departments (EDs) is a challenging task, due to high variability in arrivals and limited available resources. Early prediction of patients' outcomes, such as discharge or hospitalisation, represents an effective strategy to improve resource allocation and planning. In this study, we investigate the application of Large Language Models (LLMs) to tackle this challenge. In particular, this study explores a pipeline for transforming structured clinical event logs into narrative texts. Each patient trace is converted into a clinical story that embeds context such as triage severity, concurrent patient load, care shift, and day of the week. The narrative generation process is based on prompt-based interaction with LLMs, with prompts tuned to ensure semantic accuracy. These narrative texts are then encoded using pre-trained transformer models to produce fixed-length embeddings, which serve as input to classifiers. Applied to real-world ED data collected over a one-month period, the pipeline demonstrates how narrative episodes derived from event logs can bridge process mining and language processing, enabling outcome prediction in a format that is both informative and accessible for clinical decision-making.

Keywords: Predictive Process Monitoring · Narrative Encoding · Emergency Care pathways.

1 Introduction

Managing patient flow in emergency departments (EDs) is a challenging task due to overcrowding, unpredictable patient arrivals, limited resources, and delays in transitions between care stages, often leading to long wait times and reduced care quality. A promising strategy to address this challenge consists in leveraging artificial intelligence techniques. In particular, early prediction of patients' outcomes, such as discharge or hospitalisation, can support more effective

resource allocation and planning. In recent years, Machine Learning (ML) and Deep Learning (DL) have been widely applied to healthcare prediction tasks [2], with a growing trend towards using LLMs for outcome prediction. Despite their strong generalisation across domains, their use on structured healthcare data is still in its early stages [31].

This study aims to contribute to this emerging field by providing a real case study showcasing the application of LLMs for outcome prediction in ED processes leveraging process data. Rather than relying solely on structured event logs, we adopt a narrative-based view of patient histories, aiming to retain temporal structure and contextual information. Narratives are encoded using pretrained transformers to generate representations suitable for classification. The approach builds on techniques from process mining, language modelling, and clinical decision support, and is evaluated on real-world data from a medium-sized Italian hospital. Finally, the generated narratives proved coherent and informative, preserving the temporal structure and key clinical elements of the original traces, as confirmed by domain expert review. The classification task has been performed with several models, enabling a comparison of their performance; all achieved solid predictive performance while maintaining acceptable computation times suitable for practical deployment. The experiment has been carried out under the supervision of domain experts (the ED director and her team) to ensure relevance and validity.

The remainder of the paper is organised as follows. Section 2 introduces the related works. In Section 3, we describe the case study, and in Section 4, we outline the methodology. Section 5 details the results achieved, Section 6 discusses the results achieved, and Section 7 concludes the paper.

2 Related work

In healthcare, Process Mining (PM) includes a set of techniques for analysing and improving healthcare processes through event log data [15], e.g. for process discovery of real-world patient flow or identifying inefficiencies from event logs of hospital information systems (HIS). The application of ML to PM has led to the emergence of Predictive Process Monitoring (PPM) as a research field to forecast the evolution of ongoing cases [9,30] with applications not only in healthcare but also, for example, in education [19,20] and law [21,16]. Initially focused on individual process instances, recent efforts have shifted toward incorporating contextual and temporal dynamics, such as resource utilisation or workload, to improve prediction accuracy [27]. With the rise of unstructured and semi-structured data, NLP techniques, particularly transformer-based solutions, have become increasingly relevant for process-oriented analysis. LLMs are widely adopted for their ability to capture semantic patterns that are often missed by traditional methods, thus enriching analysis with a deeper contextual understanding [10]. Applications include event log enrichment [18], process abstraction [4], and semantic context modeling [25,17]. Models such as BERT [8], GPT [22], and derivatives have shown promise in supporting process-oriented tasks through richer event

semantics. Two main lines of work inform our approach. Transformer-based techniques for PPM, such as ProcessTransformer [5], model event sequences for next activity and remaining time prediction. Narrative-based methods like LUPIN [24] convert event logs into textual narratives to support suffix prediction with a fine-tuned Medium BERT model.

Building on these contributions, our study shifts from sequence to outcome prediction. While LUPIN relies on supervised fine-tuning with structured input, we instead generate free-text narratives via prompt-based LLMs and apply sentence embeddings without further training.

In a recent work [12], we validated the ED process, and we tested Medium BERT. Here we explore several new research directions: first, by replacing structured input with LLM-generated clinical narratives; second, by removing the need for supervised fine-tuning, thus reducing computational costs and enabling scalable experimentation across datasets; in addition, we examine both proprietary and publicly available models based on open-source architectures.

3 Case Study

This study draws on a real-world event log collected from the ED of San Luigi Gonzaga Hospital[4], a medium-sized healthcare facility located in Orbassano, near Turin (Italy). The dataset covers one month in an ED with approximately 3,500 patient admissions, averaging 125 cases per day. It traces each patient's journey from triage to discharge, recording clinical activities such as physician assessments, laboratory tests, imaging procedures, and consultations. All procedures are annotated with ICD-9 codes to ensure standardisation and interpretability. The validation of the dataset has been addressed in [12].

Each activity includes both start and end timestamps, enabling a detailed reconstruction of care pathways and service durations. At triage, patients are assigned an Emergency Severity Index (ESI) score, ranging from 1 (most urgent) to 5 (least urgent), which informs case prioritisation during the ED stay. Hospital managers and clinicians need to distinguish between cases requiring admission and those suitable for discharge. The outcome of each case was labelled as discharged or hospitalised, providing a binary target for modelling. Domain experts involved were the ED director, physicians, managers, and nurses.

4 Methodology

This Section outlines the methodology to transform a clinical event log into narrative text for outcome prediction. Figure 1 shows the pipeline from data extraction and enrichment to narrative generation, embedding, and classification. Details on log extraction and enrichment are reported in [12].

Dataset Extraction. From the HIS, we extracted the information needed to generate a timed event log for process-oriented analysis. Each record includes a

[4] Orbassano San Luigi Hospital, https://www.sanluigi.piemonte.it/web/it

Fig. 1. Methodological pipeline adopted in this research, from data extraction and event log enhancement to narrative generation, sentence embedding, and final classification. Full-size images are publicly available: https://github.com/roberto-nai/PODS4H2025.

patient identifier, the clinical treatment performed, its timestamp, the ESI level, the operator's identifier, and the outcome of the care trajectory. All treatments carried out during the ED stay (e.g., triage, physician assessments, laboratory tests, consultations) are coded using the ICD-9 taxonomy.

Event log construction. Following standard event log structure guidelines [1], we used the randomly assigned patient identifier from the HIS as the *Case ID*, the clinical treatment performed as the *activity*, and the original timestamp recorded for each procedure as the *timestamp*. The resulting event log has been exported in CSV format and tested using process discovery techniques and discussed with domain experts to validate its structure, temporal consistency, and case integrity [12].

Event log enrichment. The event log was further enriched with contextual information to support a more detailed understanding of process dynamics and workload conditions. In particular, according to the guidance of medical staff, the following features have been added based on the activity timestamp: (i) *concurrent ESI patients*, to account for crowding and case severity at the time of each activity; (ii) *day of the week*, to capture weekly patterns in patient flow and resource allocation; (iii) *work shift* (morning, afternoon, night), to reflect differences in staff presence and clinical routines across time windows.

Event log to narrative with LLMs. To facilitate the application of language models and improve the interpretability of patient traces, structured event logs were transformed into narrative form. Each patient's *prefix* trace was converted into a coherent clinical narrative through prompt-based generation. Since prediction requires observing only a partial sequence of the process, we define a prefix trace as the sequence of events available before a prediction point within a patient's ED stay. To reflect realistic decision-making scenarios, we focused on prefixes covering up to half of each case. This is a choice supported by previous studies showing that mid-case prefixes offer a good trade-off between early intervention and predictive performance [29].

After defining the prefix length, we designed structured prompts to generate semantically consistent narratives that preserved temporal ordering and key clinical features. We experimented with two types of prompts: free-form, which allows flexible and natural language generation, and template-guided, which follow predefined structures to ensure consistency. Given the need for control and alignment with the event log, we primarily adopted template-guided prompts.

This approach builds on recent studies exploring the use of LLMs to verbalise structured logs for human interpretation and AI tasks [23]. Figure 2 illustrates the prompt formulations considered in this study: two free-form (Prompts 1 and 2) and one template-guided (Prompt 3). Each prompt introduces increasing levels of structure, from loosely defined instructions to a highly constrained format with explicit placeholders. The template-guided prompt was ultimately adopted for the experiment, as it provided better control over content structure and ensured alignment with the underlying event log semantics. The pipeline supports multiple LLMs, and we selected proprietary models from the OpenAI GPT family, specifically GPT-4. The dataset was partitioned to support prompt refinement and subsequent evaluation of output consistency and semantic fidelity. As mentioned, domain experts overviewed the alignment of the generated texts with the underlying clinical facts. Once the best prompt was selected, the full dataset was processed [28]. Prompt-based generation was preferred over traditional pipelines, as it reduces the need for manual feature engineering and enhances adaptability across datasets. Finally, to reduce execution time, narrative generation was parallelised using a thread-based approach.

```
# Prompt 1
[
  {"role": "system",
   "content": (
     "You are a clinical assistant. Given a patient's emergency department visit described as a sequence of events, "
     "write a short narrative that describes what happened during the visit in natural and medically accurate language."
   )
  },
  {"role": "user", "content": event_log}
]
# Prompt 2
[
  {"role": "system",
   "content": (
     "You are a documentation model trained on emergency care data. "
     "Given a list of clinical events from a patient trace, generate a narrative that could appear in a medical report. "
     "Begin the narrative with: 'The patient arrived at the emergency department and...'"
   )
  },
  {"role": "user", "content": event_log}
]
# Prompt 3
[
  {"role": "system",
   "content": (
     "You are a clinical report generator. Based on the following structured emergency department events, "
     "produce a narrative following this format:\n\n"
     "The day {{DAY}}, {{RESOURCE}} started performing {{ACTIVITY}} at {{TIMESTAMP}} on a patient. "
     "The patient was assigned with ESI {{ESI}}. "
     "At the moment of the activity, {{CURRENT_ESI_1}} patients had a ESI of level 1, "
     "{{CURRENT_ESI_2}} patients had a ESI of level 2, {{CURRENT_ESI_3}} patients had a ESI of level 3, "
     "{{CURRENT_ESI_4}} patients had a ESI of level 4, {{CURRENT_ESI_5}} patients had a ESI of level 5. "
     "The activity was performed during the {{SHIFT}} shift. "
     "The outcome for the patient is {{OUTCOME}}.\n\n"
     "Fill in the placeholders with values extracted from the data and write fluent English sentences."
   )
  },
  {"role": "user", "content": event_log}
]
```

Fig. 2. Prompt formulations used in the study. Each prompt introduces increasing structure, from the open-ended task in Prompt 1 to the guided sentence opening in Prompt 2, up to the fully constrained template with explicit labels in Prompt 3, designed to steer LLM behaviour during narrative generation

Classification via sentence embeddings. The generated narratives were encoded using pre-trained transformer-based models available in the Sentence-

Transformer (ST) library [32]. We experimented with eight publicly available models (architecture and developer indicated in parentheses): `all-mpnet-base-v2` (MPNet, Microsoft), `intfloat/e5-base-v2` (BERT, Hugging Face), `BioBERT-mnli-snli-scinli-scitail-mednli-stsb` (BERT, DMIS Korea), `pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb` (BERT, DMIS Korea & Pritam Deka), `nli-roberta-base-v2` (RoBERTa, Facebook AI), `multi-qa-mpnet-base-dot-v1` (MPNet, Microsoft), `sentence-t5-base` (T5, Google), and `princeton-nlp/sup-simcse-roberta-base` (RoBERTa, Princeton NLP). These models produce fixed-length dense vector representations [26] of clinical narratives, capturing both semantic and contextual information in a format suitable for downstream ML tasks. All models were selected for their architectural diversity (BERT, MPNet, RoBERTa, T5), domain-specific optimisations (classification, inference, biomedical), and consistent performance across standard evaluation benchmarks such as MTEB [14] and SentEval [7].

Rather than fine-tuning full transformer models, we adopted a hybrid strategy in which sentence embeddings are computed once and then used as input features for downstream classifiers. This approach leverages the effectiveness of models such as Random Forest (RF), XGBoost (XGB), and Neural Networks (NNs) trained on frozen embeddings [11]. It offers a scalable alternative to fine-tuning by reducing computational demands and enabling rapid experimentation, particularly in resource-constrained scenarios. Accordingly, the encoded narratives were used as input to three classifiers: RF, XGB, and FFNN; the latter selected for its ability to model non-linear patterns in embedding spaces while maintaining lower computational complexity compared to deeper architectures such as recurrent networks [6].

Finally, transformer encoders were used in a frozen manner to generate sentence embeddings for the classifiers. A 70/15/15 train/validation/test split preserved outcome distribution, while the test set was balanced (50% per outcome). Evaluation metrics were computed: accuracy, precision, recall, and F1-score.

Implementation details. The experiments were conducted in Python 3.12 (64-bit). Model calls used the official APIs of OpenAI (GPT-4o[5]) and Hugging Face for ST[6]. The FFNN was implemented in PyTorch (v2.2.2) with three dense layers, LeakyReLU, dropout, and batch normalisation, while RF and XGB relied on scikit-learn (v1.7). Hyperparameters were optimised with Hyperopt [3]. All scripts are publicly available[7].

5 Results

This Section presents the results of our approach: description of the event log, its transformation into narratives, and the evaluation of outcome prediction across

[5] https://platform.openai.com/docs/models/gpt-4o

[6] https://huggingface.co/sentence-transformers

[7] https://github.com/roberto-nai/PODS4H-2025 – the event log is not public due to privacy constraints; contact the corresponding author for access.

embedding models and classifiers. Experiments were conducted on an ARM-based system with a 3.2 GHz (10-core CPU / 24-core GPU) processor and 32 GB of RAM.

Event log description. The event log obtained includes 3,478 cases, with a high degree of variability (1,036 variants) in patients' pathways; cases have a median duration of 2.9 hours and an average duration of 9.4 hours. The dataset comprises 456 hospitalisation cases and 2,807 discharge cases, reflecting a class imbalance. Table 1 summarises the ED event log by patient outcome. Hospitalised cases show greater complexity, with longer durations and more variants, high average Entropy (ENT) and Cyclomatic Complexity (CC). Domain experts explained that these marked differences can be attributed to the clinical management of patients. Those who are eventually hospitalised often undergo a Short Stay unit (*OBI*, in Italian), which substantially extends their stay in the ED and contributes to the higher variability and complexity of their care pathways. In contrast, discharged patients typically present lower severity levels and do not require OBI, resulting in shorter and more standardised trajectories.

Table 1. Summary of the ED event log split by outcome (D = Discharged, H = Hospitalised); ENT = entropy of activity transitions; CC = Cyclomatic Complexity.

Outcome	#Traces	#Variants	#Activities	Median case duration (h)	Mean case duration (h)	Mean ENT (std)	Mean CC (std)
D	2,807	650	47	2.6	7	1.80 (0.64)	1.44 (0.90)
H	456	395	40	20.2	25.5	2.85 (0.58)	3.02 (1.85)

Event log to narrative with LLMs. The event log is converted into text, as illustrated in the following example. The event log in CSV format (Figure 3) has been transformed into a text version in JSON format (Figure 4) The text describes how a nurse (NURSE_6) performed a triage activity at a specific timestamp, including contextual information about the number of patients per ESI level at that moment. It also reports the patient's assigned ESI score and the outcome (hospitalised). The results of this phase have been inspected by domain experts and found to be consistent.

```
CASEID,ACTIVITY,TIMESTAMP,TIMESTAMP_END,RESOURCE,ESI,DATE,DAY_start,DAY_end,CURRENT_ESI_1,CURRENT_ESI_2,CURRENT_ESI_3,CURRENT_ESI_4,CURRENT_ESI_5,SHIFT,OUTCOME
2022090067,TRIAGE,2022-09-01 13:29:25,2022-09-01 13:36:00,NURS_6,1,2022-09-01,Thursday,Thursday,3,0,0,0,0,Morning,hospitalised
...
2022090067,PHYSICIAN-ASSESMENT,2022-09-01 13:36:00,2022-09-01 13:36:00,DOCT_7,1,2022-09-01,Thursday,,3,0,0,0,0,Morning,hospitalised
...
2022090067,DISCHARGE,2022-09-01 15:01:00,2022-09-01 15:01:00,-,1,2022-09-01,Thursday,,3,0,0,0,0,Evening,hospitalised
```

Fig. 3. Event log representation of a patient prefix trace used for narrative generation (in CSV format)

Classification via sentence embeddings. Table 2 presents the top 10 combinations of sentence embedding models and classifiers, ranked by F1-score. The best overall result was achieved by the FFNN using `multi-qa-mpnet-base-dot-v1`,

```
{...},
{
  "case_id": 2022090067,
  "narrative":
  "The day Thursday, NURS_6 started performing TRIAGE at 2022-09-01 13:29:25 on a patient. At the moment of
the activity, 3 patients had a ESI of level 1, 0 patients had a ESI of level 2, 0 patients had a ESI of level
3, 0 patients had a ESI of level 4, 0 patients had a ESI of level 5. The activity was performed during the
Morning shift.
    ...
    The day Thursday, DOCT_7 started performing PHYSICIAN-ASSESMENT at 2022-09-01 13:36:00 on a patient. At
the moment of the activity, 3 patients had a ESI of level 1, 0 patients had a ESI of level 2, 0 patients had
a ESI of level 3, 0 patients had a ESI of level 4, 0 patients had a ESI of level 5. The activity was
performed during the Morning shift.
    ...
    The patient was assigned with ESI 1.",
    "true_outcome": "hospitalised",
},
{...}
```


Fig. 4. Example of a prefix trace converted into narrative text (in JSON format)

with an F1-score of 0.868 and high recall (0.901), indicating strong sensitivity in detecting hospitalised cases. Other high-performing combinations include XGB paired with `sup-simcse-roberta-base` and `multi-qa-mpnet-base-dot-v1`, both of which show balanced precision and recall above 0.83. RF classifiers also performed competitively, particularly when combined with `e5-base-v2` and `all-mpnet-base-v2`.

Several models not appearing in the top 10, including `sentence-t5-base`, `pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb`, and `nli-roberta-base-v2`, showed weaker overall performance.

Table 3 aggregates results by classifier type, computing the average metrics over the best-performing 10 entries. The FFNN leads with the highest average F1-score (0.818), followed by XGB (0.808) and RF (0.805). These results confirm the robustness of the FFNN in this setting, especially when combined with high-quality sentence embeddings.

In our exploratory approach to test multiple model–classifier combinations, some patterns can be observed among the best-performing setups. For instance, `multi-qa-mpnet-base-dot-v1` and `sup-simcse-roberta-base` performed well, likely due to their training on semantic similarity and retrieval tasks, which aligns with capturing clinically meaningful patterns in the narratives. These objectives align well with our narratives, which, though brief, embed structured and meaningful content. MPNet-based models seem particularly effective at capturing such details due to their handling of sentence structure and word relationships. On the classifier side, the FFNN achieved the best overall performance, likely because it can capture complex, non-linear patterns in the sentence embeddings. This ability is advantageous in our setting, as the FFNN captures details such as triage level, timing, or workload implicitly encoded in the text, achieving a good balance between accuracy and sensitivity with high-quality embeddings.

Timing of the experiments. Narrative generation with GPT-4 took approximately 30 minutes, utilising three parallel threads. Although latency can vary with input length and server load, API interactions proved stable with no failures. Classifier execution times ranged from about 8 minutes (RF) to 14 minutes (FFNN) per model–classifier combination, which is reasonable for this task.

Table 2. Top-10 classifiers per embedding model ranked by F1-score. Best result in **bold**, second-best in *italics*.

Model	Classifier	Accuracy	Precision	Recall	F1-score
multi-qa-mpnet-base-dot-v1	**FFNN**	0.863	0.837	0.901	**0.868**
princeton-nlp/sup-simcse-roberta-base	*XGB*	0.847	0.832	0.868	*0.849*
multi-qa-mpnet-base-dot-v1	XGB	0.836	0.835	0.835	0.835
intfloat/e5-base-v2	RF	0.825	0.804	0.857	0.830
princeton-nlp/sup-simcse-roberta-base	RF	0.825	0.817	0.835	0.826
all-mpnet-base-v2	RF	0.820	0.796	0.857	0.825
multi-qa-mpnet-base-dot-v1	RF	0.825	0.831	0.813	0.822
intfloat/e5-base-v2	FFNN	0.814	0.794	0.846	0.819
intfloat/e5-base-v2	XGB	0.809	0.804	0.813	0.809
princeton-nlp/sup-simcse-roberta-base	FFNN	0.809	0.811	0.802	0.807

Table 3. Average performance by classifier model-classifier combinations. Best result in **bold**, second-best in *italics*.

Classifier	Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F1-score
FFNN	0.830	0.810	0.868	**0.834**
XGB	0.830	0.824	0.851	*0.835*
RF	0.825	0.804	0.857	0.828

6 Discussion

Discussion of the results leads to different insights. Considering the absence of personal information and data related to clinical examination parameters, the obtained results appear remarkable. Moreover, as suggested by domain experts, outcome predictions, if available in real-time, could help prioritise diagnostics, trigger early consultations and support timely decisions.

In our case study, as medical decisions regarding hospitalisation should be made within six hours of patient arrival (an established practice, according to domain experts), this study appears promising toward the development of a system capable of supporting timely clinical decision-making. The flexibility of the proposed pipeline allows for the exploration of different prefix lengths to analyse how predictive signals develop throughout the patient journey, providing insight into the trade-off between early prediction and accuracy. The ability to generate predictions from narrative descriptions of ward activities also offers a potentially useful mode of interaction between clinical staff and hospital management. Nonetheless, the appropriateness of predicted outcomes remains a key consideration, as final decisions ultimately rely on clinical judgement.

Finally, the present work does not include a direct comparison with established predictive process monitoring methods, as the main objective was to test the feasibility of narrative-based representations and their execution within reasonable computation times, which are crucial in hospital settings. While narrative generation with LLMs requires extra computation compared to standard embeddings, the trade-off is justified by the added interpretability of narratives over raw event logs.

7 Conclusion and Future Work

This study demonstrates that transforming structured clinical event logs into narrative representations supports outcome prediction through the contextual understanding of LLMs, with a pipeline that runs in a reasonable time and is suitable for hospital use.

Future research may proceed along three main directions: (i) integrating explainability techniques [13] to better understand which narrative components influence predictions, including methods that highlight the most relevant input features, such as local and global explanations or counterfactual reasoning; (ii) comparing this approach with traditional PPM methods, considering established methodologies, tools, and encoding techniques to identify those most suitable for narrative-based prediction; (iii) exploring multi-task settings and assessing generalisability across hospitals and departments to evaluate the robustness of the approach beyond the initial case study.

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.

References

1. van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer, 2nd edn. (2016). https://doi.org/10.1007/978-3-662-49851-4
2. Badawy, M., Ramadan, N., Hefny, H.A.: Healthcare predictive analytics using machine learning and deep learning techniques: a survey. Journal of Electrical Systems and Information Technology **10**(1), 40 (2023). https://doi.org/https://doi.org/10.1186/s43067-023-00108-y
3. Bergstra, J., Yamins, D., Cox, D.D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proc. of ICML. Proceedings of Machine Learning Research, vol. 28, pp. 115–123. PMLR (2013). https://doi.org/10.48550/arXiv.1206.2944
4. Berti, A., Schuster, D., van der Aalst, W.M.P.: Abstractions, scenarios, and prompt definitions for process mining with llms: A case study. In: BPM Workshops 2023. LNBIP, vol. 492, pp. 427–439. Springer (2023). https://doi.org/10.1007/978-3-031-50974-2_32
5. Bukhsh, Z., Tabatabaei, S.A., Dijkman, R., Grefen, P.: Processtransformer: Predictive business process monitoring with transformer network. Information Systems **107**, 102784 (2022). https://doi.org/10.1016/j.is.2022.102784
6. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, B., Strope, B., Kurzweil, R.: Universal sentence encoder. In: Proc. of EMNLP: System Demonstrations. pp. 169–174. ACL (2018). https://doi.org/10.18653/v1/D18-2029

7. Conneau, A., Kiela, D.: Senteval: An evaluation toolkit for universal sentence representations. In: Proceedings of LREC 2018. pp. 979–984 (2018), http://www.lrec-conf.org/proceedings/lrec2018/pdf/721.pdf
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of NAACL: HLT. pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/N19-1423
9. Di Francescomarino, C., Ghidini, C.: Predictive process monitoring. In: Process Mining Handbook, LNBIP, vol. 448, pp. 320–346. Springer (2022). https://doi.org/10.1007/978-3-031-08848-3_10
10. Estrada-Torres, B., del-Río-Ortega, A., Resinas, M.: Mapping the landscape: Exploring large language model applications in business process management. In: Enterprise, Business-Process and Information Systems Modeling Proceedings. LNBIP, vol. 511, pp. 22–31. Springer (2024). https://doi.org/10.1007/978-3-031-61007-3_3
11. Li, J., Zhao, W., Deng, P.: Hybrid text classification using transformer-based embeddings and lightweight classifiers. J. Ambient Intell. Humaniz. Comput. (2023). https://doi.org/10.1007/s12652-023-04677-5
12. Lovera Ruffi, V., Nai, R., Sulis, E., Di Caro, L., Genga, L., Boccuzzi, A.: Transforming event logs into narrative texts for outcome prediction: A case study in a hospital emergency department. In: Proceedings of the 2nd International Workshop on Process Mining Applications for Healthcare (PM4H 2025) co-located with the 23rd International Conference on Artificial Intelligence in Medicine (AIME 2025) (2025), to appear
13. Meo, R., Nai, R., Sulis, E.: Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next? In: Advances in Databases and Information Systems - 26th European Conference, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings. LNCS, vol. 13389, pp. 25–34. Springer (2022). https://doi.org/10.1007/978-3-031-15740-0_3
14. Muennighoff, N., Wang, Y., Scholz, S., Magne, L., Li, T., Reimers, N., Gurevych, I.: Mteb: Massive text embedding benchmark. In: Proc. of EMNLP. pp. 2681–2700 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.167
15. Munoz-Gama, Jorge, et al.: Process mining for healthcare: Characteristics and challenges. J. Biomed. Informatics **127**, 103994 (2022). https://doi.org/10.1016/j.jbi.2022.103994
16. Nai, R., Fatima, I., Morina, G., Sulis, E., Genga, L., Meo, R., Pasteris, P.: AI applied to the analysis of the contracts of the italian public administrations. In: 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. CEUR Workshop Proceedings, vol. 3486, pp. 255–260. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3486/100.pdf
17. Nai, R., Sulis, E., Fatima, I., Meo, R.: Large language models and recommendation systems: A proof-of-concept study on public procurements. In: NLDB 2024, Turin, Italy, June 25-27, 2024, Proceedings. LNCS, vol. 14763, pp. 280–290. Springer (2024). https://doi.org/10.1007/978-3-031-70242-6_27
18. Nai, R., Sulis, E., Genga, L.: Automated analysis with event log enrichment of the european public procurement processes. In: Proc. of Advances in Conceptual Modeling Workshops. LNCS, vol. 14319, pp. 178–188. Springer (2023). https://doi.org/10.1007/978-3-031-47112-4_17
19. Nai, R., Sulis, E., Genga, L.: Enhancing e-learning effectiveness: a process mining approach for short-term tutorials. J. Intell. Inf. Syst. **62**(6), 1773–1794 (2024). https://doi.org/10.1007/S10844-024-00874-9

20. Nai, R., Sulis, E., Marengo, E., Vinai, M., Capecchi, S.: Process mining on students' web learning traces: A case study with an ethnographic analysis. In: Responsive and Sustainable Educational Futures - 18th EC-TEL 2023, Aveiro, Portugal, September 4-8, 2023, Proceedings. LNCS, vol. 14200, pp. 599–604. Springer (2023). https://doi.org/10.1007/978-3-031-42682-7_48
21. Nai, R., Sulis, E., Pasteris, P., Giunta, M., Meo, R.: 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. Communications in Computer and Information Science, vol. 1752, pp. 89–102. Springer (2022). https://doi.org/10.1007/978-3-031-23618-1_6
22. OpenAI: Gpt-4 technical report (2023). https://doi.org/10.48550/arXiv.2303.08774
23. Papadopoulos, S., Symeonidis, C., Tsinaraki, C.: Transforming event logs into narratives with large language models: A feasibility study. In: BPM Workshops. pp. 3–15. Springer (2023). https://doi.org/10.1007/978-3-031-37808-0_1
24. Pasquadibisceglie, V., Appice, A., Malerba, D.: Lupin: A llm approach for activity suffix prediction in business process event logs. In: 2024 6th International Conference on Process Mining (ICPM). pp. 1–8 (2024). https://doi.org/10.1109/ICPM63005.2024.10680620
25. Rebmann, A., Schmidt, F.D., Glavaš, G., van Der Aa, H.: Evaluating the ability of llms to solve semantics-aware process mining tasks. In: 2024 6th International Conference on Process Mining (ICPM). pp. 9–16 (2024). https://doi.org/10.1109/ICPM63005.2024.10680677
26. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proc. of EMNLP 2019. pp. 3982–3992 (2019). https://doi.org/10.18653/v1/D19-1410
27. Senderovich, A., Di Francescomarino, C., Ghidini, C., Jorbina, K., Maggi, F.M.: Intra and inter-case features in predictive process monitoring: A tale of two dimensions. In: Proc. of BPM 2017. LNCS, vol. 10445, pp. 306–323. Springer (2017). https://doi.org/10.1007/978-3-319-65000-5_18
28. Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., Wang, Y.: An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. arXiv preprint arXiv:2309.08008 (2023). https://doi.org/10.48550/arXiv.2309.08008
29. Teinemaa, I., Dumas, M., La Rosa, M., Maggi, F.M.: Outcome-oriented predictive process monitoring: Review and benchmark. ACM Trans. Knowl. Discov. Data (TKDD) **13**(2), 1–57 (2019). https://doi.org/10.1145/3297750
30. Verenich, I., Dumas, M., Rosa, M.L., Maggi, F.M., Teinemaa, I.: Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. ACM Trans. Intell. Syst. Technol. **10**(4), 34:1–34:34 (2019). https://doi.org/10.1145/3331449
31. Wang, Y., Zhang, Y., Lin, Z., Wang, Y., Xie, X., Chen, E.: Large language models for healthcare: A survey. Patterns **4**(10), 100791 (2023). https://doi.org/10.1016/j.patter.2023.100791
32. Wolf, T., Debut, L., Sanh, V., Chaumond, J., et al.: Transformers: State-of-the-art natural language processing. In: Proc. of EMNLP 2020: System Demonstrations (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6