ARTICLE TITLE: **SPELL-LLMs: A Scalable and Privacy-Compliant NLP Pipeline Using Locally Hosted Large Language Models for Clinical Information Extraction**

Author information: Ricardo Kleinlein, Ph.D.[1], David W. Bates MD MSc[2,3], Kathryn J. Gray, MD PhD[4*], Vesela P. Kovacheva MD PhD[1*]

[1]Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[2]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA

[3]Department of Health Care Policy and Management, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[4]Department of Obstetrics and Gynecology, University of Washington, Seattle, WA, USA

***Corresponding Author:**

Vesela Kovacheva MD PhD

75 Francis St; L1, Boston, MA 02115

Office: 617-732-8220

Fax: 617-730-2825

Email: vkovacheva@bwh.harvard.edu

Prior Presentations: None

Word and key-element counts:

Number of Figures: 3

Number of Tables: 4

Number of Appendices: 1

Word count: up to 4,000 words: 3621

Word count, excluding title page, abstract, references, figures, and tables

Abbreviated title: Scalable, Privacy-Compliant NLP Pipeline (SPELL-LLMs).

Summary Statement: We developed SPELL-LLMs, a scalable, privacy-compliant NLP pipeline combining regex and locally hosted LLMs for efficient information extraction from clinical narratives.

Keywords (up to 5): Electronic health records, Large Language Models, regular expressions, phenotyping, privacy-compliant natural language processing

Abbreviations:

BWH, Brigham and Women's Hospital

EHR, Electronic health records

MGB, Mass General Brigham

MGH, Massachusetts General Hospital

NWH, Newton Wellesley Hospital

WDH, Wentworth-Douglas Hospital

SRH, Spaulding Rehabilitation Hospital

FH, Faulkner's Hospital

NSM, North Shore Medical Center

MEE, Mass Eye and Ear

LLM, Large Language Model

ABSTRACT

Objective:

Electronic health records (EHRs) contain valuable information for clinical research and decision-making. However, leveraging these data remains challenging due to data heterogeneity, inconsistent documentation, missing information, and evolving terminology, especially within unstructured clinical notes. We developed a scalable, privacy-preserving natural language processing (NLP) workflow to systematically extract structured clinical insights from large volumes of clinical narratives.

Materials and Methods:

Our platform employs a hybrid approach combining regular expressions (regex) to rapidly identify relevant textual snippets with locally hosted large language models (LLMs) for accurate clinical interpretation. All data processing occurs securely within institutional environments, adhering strictly to data privacy regulations. The modular Python-based workflow facilitates adaptation across institutions and is optimized for computational efficiency, supporting high-throughput processing even in resource-limited settings.

Results:

The pipeline efficiently processed millions of clinical reports (1976–2024) from multiple hospitals. By analyzing targeted snippets rather than entire documents, our approach reduced processing time by 80% compared to traditional full-document LLM inference, and by 97% compared to manual physician annotation. Accuracy was rigorously validated using three obstetric tasks: extraction of numerical values (blood loss volumes), dates (estimated delivery dates), and diagnoses (hemolysis, elevated liver enzymes, and low platelets [HELLP] syndrome), achieving 95% agreement with expert annotations. Generalizability was further

confirmed by accurately identifying ventricular tachycardia diagnoses in the publicly available MT Samples dataset.

Discussion and Conclusions:

Our hybrid NLP framework significantly enhances the usability of unstructured EHR data for clinical research, decision support, and large-scale retrospective analyses.

INTRODUCTION

Electronic health records (EHRs) contain extensive clinical data in structured forms (e.g., vital signs, lab results, medications) and unstructured forms such as physician notes and discharge summaries. Integrating these data sources can enhance clinical decision-making,[1] retrospective research,[2,3] and predictive analytics,[4,5] yet leveraging clinical text remains challenging due to data heterogeneity, inconsistent documentation, and the complexity inherent in clinical narratives. [6,7]

Structured EHR data can be readily analyzed with statistical or computational methods, whereas unstructured notes often contain essential clinical details not captured elsewhere.[5] Manual annotation of clinical text is labor-intensive, [8,9] resource-consuming,[9,10] and subject to human variability, [11] limiting large-scale or longitudinal studies.[10] Therefore, automated natural language processing (NLP) solutions are crucial.

Historically, clinical NLP has relied heavily on rule-based and statistical methods,[12] which require extensive manual feature engineering and frequent retraining.[8] Recently, large language models (LLMs), capable of zero- or few-shot inference,[13] have emerged as promising tools for clinical text extraction, significantly reducing annotation effort and enhancing contextual accuracy.[13–15] However, most current LLM solutions rely on cloud-based systems, raising substantial privacy and compliance concerns due to external processing of protected health information.[13]

To address these limitations, we developed SPELL-LLMs, a scalable, privacy-preserving NLP pipeline combining regular expressions (regex) for efficient snippet extraction with locally hosted LLM inference. By restricting LLM inference to targeted snippets identified via regex, we

significantly reduce computational costs, latency, and privacy risks associated with external processing.

We demonstrate the clinical utility and scalability of SPELL-LLMs through three obstetric use cases: quantifying delivery-related blood loss, extracting estimated due dates (EDD), and identifying diagnoses or clinical suspicion of hemolysis, elevated liver enzymes, and low platelets (HELLP) syndrome. These tasks were selected due to their clinical relevance, reliance on narrative data, and limited structured documentation. To further assess generalizability and validate the flexibility of our NLP pipeline beyond obstetric scenarios, we additionally evaluated SPELL-LLMs on identifying ventricular tachycardia diagnoses using the publicly available MT Samples dataset, comprising diverse clinical specialties.

This manuscript describes our methodology, systematically evaluates accuracy, efficiency, and scalability, and discusses practical considerations for clinical deployment.

OBJECTIVE

The primary objective of this research is to introduce and validate SPELL-LLMs, a scalable, privacy-preserving clinical NLP pipeline designed to efficiently extract structured clinical information from unstructured EHR narratives. Specifically, we aim to describe a systematic workflow for retrieving and preprocessing clinical text data from institutional repositories, demonstrate the feasibility of integrating regex-based snippet retrieval methods with locally hosted LLMs for automated clinical information extraction, quantify computational efficiency and clinician-time savings compared to manual annotation methods, and validate practical utility through obstetric-related clinical use cases. Ultimately, we aim to deliver a flexible and scalable NLP solution that enhances the usability of routine clinical data for retrospective research, clinical decision support, and predictive analytics.

BACKGROUND AND SIGNIFICANCE

Early clinical NLP systems relied on manually crafted, rule-based methods, which were effective primarily for numeric data and structured entities, but were labor-intensive and lacked flexibility for diverse clinical documentation. [1,12,16,17]

The emergence of transformer-based language models (e.g., BioBERT, ClinicalBERT) significantly advanced clinical text analysis through enhanced context-aware interpretation.[18–21] However, these models require extensive annotated data and domain-specific fine-tuning, limiting practical scalability in clinical workflows.[18]

Generative large language models (LLMs), such as GPT-family models, enable efficient zero- or few-shot information extraction, outperforming previous NLP methods.[14,15,22,23] However, their computational requirements typically necessitate cloud infrastructure,[24–26] raising data privacy and compliance concerns in healthcare settings.[27]

Current clinical NLP systems, such as cTAKES and MetaMap,[28,29] utilize standard terminologies (e.g., UMLS, SNOMED-CT, ICD), but they heavily depend on vocabulary coverage, leading to incomplete extractions for rare or context-dependent concepts.[3] These tools also require separate modules to handle modifiers (such as negation and temporality), introducing additional complexity.[30]

Recent studies have combined regex-based preprocessing with generative LLMs for clinical data extraction. For example, one recent study integrated regular expressions with GPT-4 to extract surgical data but relied on cloud-based infrastructure.[31] Another evaluation of open-

source LLMs for extracting structured social determinants of health (SDoH) data demonstrated improved performance compared to traditional regex methods; however, that study did not address privacy considerations or emphasize local deployment.[15] Similarly, a clinical entity-augmented retrieval method demonstrated computational efficiency gains by leveraging clinical entities for information retrieval, yet did not prioritize local model deployment or explicit privacy compliance.[32]

SPELL-LLMs integrates regex-based preprocessing and a locally hosted generative LLM, combining computational efficiency and accurate contextual interpretation while ensuring data privacy compliance. Compared to existing methods (Table 1), SPELL-LLMs notably reduces annotation requirements, enhances flexibility, implicitly manages linguistic modifiers, and facilitates scalable deployment in privacy-sensitive healthcare environments.

Table 1. Comparison of Clinical NLP Methodologies: SPELL-LLMs versus Traditional Approaches

Criteria	**SPELL-LLM**	**NER-based (ClinicalBERT)**	**cTAKES / MetaMap**
Annotation Burden	Low	High	Medium
Flexibility	High	Medium	Medium
Computational Efficiency	High	Medium/Low	Medium
Handling Contextual Modifiers	Implicit	Requires explicit modules	Explicit
Integration with Clinical Vocabularies	Not yet implemented	Possible but requires additional mapping	Built-in
Privacy (Local hosting)	Yes	Yes (if locally hosted)	Yes (local)

Abbreviations: NLP, natural language processing

MATERIALS AND METHODS

Ethics and Data Governance

All research activities strictly adhered to institutional policies governing data privacy, ethical standards, and data governance, and received explicit approval by the Mass General Brigham Institutional Review Board (IRB#2020P002859). To ensure compliance with patient privacy regulations, all clinical data were processed locally within a secure institutional environment. Data were handled exclusively in pseudonymized form, with personally identifiable information such as patient names, medical record numbers, Social Security Numbers, contact details, and precise dates systematically obfuscated or entirely removed.

Data Sources and Dataset Generation

Clinical data used in this study were obtained from a centralized clinical data warehouse that aggregates EHR data from multiple healthcare institutions within our large network.[33] These institutions included tertiary care hospitals, community hospitals, specialty care centers, and primary care facilities.

We implemented a structured workflow for generating specialized datasets tailored for clinical NLP tasks (Figure 1). Clinical notes were provided as plain-text files. Each file contained multiple clinical reports delimited by standardized header metadata and end-of-report tokens. To facilitate efficient data organization and retrieval, custom metadata extraction scripts written in Python identified and indexed each clinical note by patient pseudonym, encounter identifier, document type, creation timestamp, and provider information. This indexing strategy improved retrieval efficiency by storing byte offsets and document lengths rather than duplicating notes, which significantly reduced redundancy.

Patient cohorts were defined flexibly, based on demographic criteria, specific clinical conditions, temporal constraints (such as gestational periods), or combinations of these factors. Cohort filtering involved explicit searches conducted within both report metadata and note text using collaboratively defined regular expressions, allowing precise selection of clinical notes relevant to clinical scenarios (Table 2).

Table 2: Examples of filters implemented in the dataset generation process.

Filter	Criterion	Examples	Inclusion
Cohort selection (predefined)	Patient IDs: [00001, 01234]	**00001**|MGH|EX-0|12/31/2014|Follow-up Visit	✓
		11111|MGH|EX-4|05/24/1992|Rheumatology Consultation	✗
Date	After 2015	01234|BWH|EX-2|**08/31/2024**|Discharge Summary	✓
		00001|MGH|EX-0|**12/31/2014**|Follow-up Visit	✗
Note type	Discharge Notes only	00042|MGH|EX-11|01/31/2018|**Discharge Summary**	✓
		11111|MGH|EX-4|05/24/1992|**Rheumatology Consultation**	✗
		00042|MGH|EX-11|01/31/2018|Discharge Summary	✓
Pregnancy #	Nulliparous	[**Assigned to patient's first pregnancy**]	
		00042|MGH|EX-1|01/31/2024|Discharge Summary	✗

		[**Assigned to patient's second pregnancy**] 01234|BWH|EX-2|02/31/2014|Discharge Summary	
Regular Expression	"(?i)\bHELLP\b"	… stage for NRFHT iso TOLAC IOL for c/f hx PPCM vs TACO iso pulm edema after first delivery c/b **HELLP**. C/b deep right uterine extension and intraop fever. QBL 1159 cc. S/p pitocin x 3 … 06789|NWH|EX-3|03/31/2014|Discharge Summary	✔
		… oblem List Diagnosis ?Low-lying placenta ?CP cysts ?SGA (small for gestational age) ? **HELLP** syndrome ?Normal pregnancy Surgical (OR) Procedures: Surgical/Procedural Cases on this Adm…	✔

Clinical Data Extraction Workflow

The data extraction process involved two sequential stages (Figure 1): (1) cohort identification and snippet extraction using regular expressions, and (2) structured clinical information retrieval using a locally hosted large language model (LLM). Figure 2 illustrates an example workflow for the extraction process.

Cohort Identification and Contextual Snippet Extraction

Clinical notes relevant to specific research questions were identified using regular expressions collaboratively defined by clinical domain experts and data scientists. These regular expressions targeted specific domain-relevant terminology (e.g., “estimated blood loss,” “due date,” “HELLP syndrome”) and extracted snippets surrounding matched terms. The search parameters—including regular expressions, snippet size, and context selection rules—were fully configurable within the extraction pipeline, providing flexibility to accommodate diverse clinical extraction scenarios. For instance, retrieving a specific date typically required minimal surrounding text, whereas discerning condition relevance (current versus past pregnancies) required broader snippets.

Extracted snippets were systematically organized within dedicated directory structures, enhancing reproducibility and enabling efficient subsequent retrieval. Additionally, the platform offered random sampling functionality for manual review of regex-based snippet extraction, facilitating either prompt validation or further refinement of regular expression patterns. Researchers could explicitly define sample sizes, enabling iterative improvements in extraction precision.

Large Language Model (LLM)-Assisted Extraction

To enhance automated clinical information extraction from the snippets, our workflow incorporated a locally deployed, privacy-compliant LLM. We selected Llama 3.1-8B-Instruct as our default model based on a combination of its architectural, functional, and deployment characteristics. Llama 3.1 is an autoregressive language model built upon an optimized transformer architecture, with the Instruct variant specifically fine-tuned for human-aligned dialogue using supervised training and reinforcement learning with human feedback. The 8-billion-parameter model was chosen because it represented the largest variant feasible for complete local deployment, thereby maintaining stringent patient privacy. Its large, supported

context window (up to 128,000 tokens) was critical for effectively processing lengthy clinical narratives without truncation, an advantage compared to medically tuned models (e.g., ClinicalBERT, BioGPT, ClinicalGPT), which have smaller context windows (typically ≤ 512 tokens) and thus limit comprehensive context analysis. Furthermore, instruction-tuned generative models often achieve superior clinical extraction accuracy compared to specialized domain-specific models.[34] The model integration was explicitly designed to be modular and adaptable to future advancements in LLM technology, allowing seamless substitution of alternative models.

The platform offered extensive flexibility in inference configuration, allowing researchers to customize essential parameters—such as temperature settings, response token limits, and prompt templates—to optimize extraction accuracy and computational efficiency. To address computational bottlenecks and enhance model effectiveness, LLM inference was restricted to contextually relevant snippets identified by regular expressions (regex), rather than processing entire clinical documents. This approach dramatically reduced runtime and token usage, enabling efficient use of smaller LLM architectures by providing precise context, thus minimizing irrelevant noise. For longer snippets exceeding fixed token limits, a configurable overlap-based segmentation strategy (50% overlap) maintained contextual continuity.

Prompts were explicitly designed for specific clinical extraction tasks, providing clear, task-oriented instructions for information retrieval without the necessity for additional fine-tuning or extensive prior prompt training (Appendix S5). Model inference settings, execution scripts, and prompt details were systematically documented for reproducibility and iterative refinement.

Hardware and Software Infrastructure

The pipeline was developed and executed on a single Linux workstation (Ubuntu 22.04 LTS) equipped with three NVIDIA RTX A4000 GPUs, each with 16 GB of VRAM. Computational environments were managed using a dedicated Miniforge installation (Python 3.13.5, CUDA 12.7), with essential computational libraries including Polars (version 1.17.1) for efficient tabular data manipulation and Transformers (version 4.45.2) for local LLM inference. The pipeline architecture emphasized scalability, modularity, computational efficiency, and reproducibility. Each workflow component—metadata indexing, regex snippet retrieval, and LLM inference—was encapsulated into clearly defined, configurable Python scripts to facilitate easy adaptation to new clinical tasks and reduce technical barriers for clinical informatics practitioners.

Clinical Use-Case Evaluations

We rigorously evaluated the practical utility and performance of the extraction pipeline across three obstetric clinical tasks representing distinct clinical extraction scenarios:

- **Blood Loss (BL):** Quantitative extraction of blood loss values documented in delivery discharge summaries.
- **Estimated Due Date (EDD):** Extraction of clinically determined estimated due dates from pregnancy-related documentation.
- **HELLP Syndrome Detection:** Binary classification of explicit diagnoses or documented suspicion of HELLP syndrome in discharge notes.

Performance metrics—including accuracy, precision, recall, mean absolute error (MAE), and root mean squared error (RMSE)—were benchmarked against expert physician annotations. As a baseline, regex-only performance was assessed using Regex@N (proportion of correct extractions within N tokens after regex matches). Computational efficiency metrics, GPU utilization statistics (temperature, power consumption), and runtime scalability details were systematically collected and documented (Appendix S1).

To evaluate SPELL-LLMs' adaptability beyond obstetrics, we additionally assessed its performance in identifying ventricular tachycardia using the publicly available MT Samples dataset (containing 5,013 notes across 40 clinical specialties).[35] Ventricular tachycardia was chosen for its clinical relevance and diverse documentation patterns. We quantified accuracy, precision, recall, F1-score, and computational efficiency to confirm the pipeline's broader applicability.

RESULTS

Dataset Overview and Retrieval

Since deployment, our extraction pipeline has successfully processed more than 14 large-scale data requests. The initial dataset retrieved from our institutional data warehouse encompassed 149 GB of clinical text, comprising 30,888,929 reports from 242,413 unique patients across eight hospitals within a large healthcare network, covering the years 1976–2024 (Appendices S2–S5). Metadata indexing, completed in 85 minutes (5,095 seconds), substantially optimized downstream querying and storage efficiency (structured metadata: 3.15 GB).

We illustrate the pipeline's accuracy, scalability, and clinical utility through three obstetric use cases: quantifying blood loss (BL), extracting estimated due dates (EDD), and detecting HELLP syndrome.

Clinical Note Selection via Regular Expressions

Clinical notes relevant to our obstetric tasks were selected using regular expressions collaboratively designed by clinical and informatics experts (Appendix S6). Targeted snippets

were extracted with a ±100-character context window around matched terms. The regex-based approach efficiently identified:

- **Blood Loss:** 1,286,161 delivery discharge notes (~104 minutes).
- **EDD:** 2,066,023 history & physical notes (~81 minutes).
- **HELLP:** 46,953 discharge notes (~70 minutes).

From these initial searches, subsets were selected for detailed downstream validation: 92,380 notes (BL), 35,172 notes (EDD), and 540 notes (HELLP). The criteria for these selections are detailed in Supplemental Material 4.

Efficiency of LLM-Assisted Data Extraction

The efficiency of our snippet-based LLM approach (Llama 3.1-8B-Instruct) was benchmarked against manual annotations by clinicians and full-document LLM inference (Table 3). Snippet-based inference provided a substantial efficiency gain (~97% relative speedup) compared to manual physician annotation, significantly reducing clinician workload.

Table 3: Annotation times (in seconds) comparing physician with LLM extraction. Standard deviations (in parentheses) are reported only for LLM annotations across 100 repetitions.

Use case (time in s)	Physician	LLM (snippets)	LLM (full report)
Blood loss	449	12.65 (0.45)	73.71 (0.36)
EDD	551	19.54 (0.16)	66.27 (0.17)
HELLP	480	10.59 (0.16)	79.96 (0.35)

Abbreviations: EDD, estimated due date; HELLP, hemolysis, elevated liver enzymes, low platelets, LLM large language model.

Efficiency was further improved through sequential dynamic loading of snippets, bypassing GPU padding bottlenecks caused by highly variable snippet lengths (12–2,000 tokens). Compared to extensively optimized traditional batch-processing methods, dynamic loading improved inference speeds by up to 86% (Figure 3).

Accuracy and Performance Evaluation

The accuracy of the LLM-assisted extraction was rigorously assessed against manual annotations (Table 4). For blood loss quantification, the LLM achieved 98% accuracy (mean absolute error [MAE]: 14 mL, root mean squared error [RMSE]: 99 mL). For EDD extraction, the LLM attained 94% accuracy (MAE: 41.5 days), with significant extraction errors (>7 days discrepancy) observed in only 2 of 50 cases, primarily due to ambiguity from multiple dates. The HELLP syndrome detection task yielded 94% accuracy, 96% precision, 98% recall, and an F1-score of 97%. Errors typically resulted from misinterpretation of the pregnancy context (current versus prior pregnancies).

By comparison, regex-only approaches demonstrated significantly lower reliability, underscoring the importance of the contextual interpretation capabilities of the LLM.

Table 4. Performance metrics comparing regex-only methods (Regex@N) versus LLM extraction.

	Metric	**Regex@1**	**Regex@3**	**Regex@5**	**LLM**
HELLP	Accuracy	0.94	0.94	0.94	0.94
	Precision	0.94	0.94	0.94	0.96
	Recall	1	1	1	0.98
	F1-Score	0.97	0.97	0.97	0.97

	No regex match found	0	0	0	-
EDD	Accuracy	0.71	0.59	0.57	0.94
	MAE (days)	443.71	646.31	646.31	41.49
	Mistaken samples	4	12	12	2
	Mistakes > 7 days	3	11	11	2
	Max error (days)	2557	2923	2923	1826
	No regex match found	36	21	20	-
Blood Loss	Accuracy	0.8	0.89	0.87	0.98
	MAE (mL)	59	39.89	65.13	14
	RMSE (mL)	186.81	153.59	221.88	98.99
	No regex match found	35	13	12	-

Abbreviations: EDD, estimated due date; HELLP, hemolysis, elevated liver enzymes, low platelets; MAE, mean absolute error; RMSE, root mean square error.

Regex@N denotes the proportion of correct extractions within N tokens following a regex match.

Computational Performance

Inference time scaled linearly with input length, demonstrating efficient computational resource use. GPU monitoring indicated stable, sustainable, and scalable hardware performance, making it suitable for institutional implementation (detailed GPU metrics are provided in Appendix S1).

Practical Workflow Example (EDD Extraction)

Figure 2 demonstrates a practical, step-by-step workflow example: regex-based note selection, snippet extraction, and structured data extraction via LLM inference. This illustrates the pipeline's potential for intuitive integration into routine clinical informatics workflows (detailed description in Appendices S7–S9).

Generalizability Evaluation Across Clinical Contexts

To evaluate SPELL-LLMs' generalizability beyond obstetrics, we tested its performance in identifying ventricular tachycardia diagnoses in the MT Samples dataset. SPELL-LLMs achieved high extraction accuracy (84%) and an F1-score of 67%, demonstrating robust performance despite substantial differences from obstetric documentation. The snippet-based inference strategy significantly enhanced efficiency, completing the task in 2.74 seconds compared to 15 minutes for full-document analysis, and yielded approximately 3.5-fold fewer false positives. These findings underscore SPELL-LLMs' adaptability and potential for broad clinical informatics applications.

DISCUSSION

In this study, we developed and validated SPELL-LLMs, a scalable, modular, and privacy-compliant NLP pipeline designed to automatically extract structured clinical insights from unstructured electronic health record (EHR) narratives. The hybrid framework combined precise regex-based snippet extraction with locally hosted large language model (LLM) inference, demonstrating substantial computational efficiency gains and high accuracy across three representative obstetric tasks: quantifying blood loss, retrieving estimated due dates(EDD), and detecting HELLP syndrome. These results were further validated using both a perioperative dataset and a publicly available dataset.

Key findings and implications

Our results demonstrated significant advantages of SPELL-LLMs compared to manual physician annotation and purely regex-based methods. Specifically, SPELL-LLMs significantly reduced manual review effort, achieving 97% time savings compared to manual annotations. Furthermore, the platform consistently matched expert-level accuracy (~95%) across diverse clinical extraction tasks, including validation with the MT Samples dataset, highlighting its potential to transform the feasibility of large-scale, retrospective clinical studies and routine clinical analytics. [2,5,13]

A critical contribution of our approach was its structured workflow, modular architecture, and inherent flexibility to integrate emerging generative AI technologies seamlessly.[13,14] Unlike specialized NLP pipelines that typically require intensive manual annotation, feature engineering, and frequent model retraining,[8,9,12] our platform was explicitly designed for rapid adaptability and ease of use. The modular strategy specifically anticipated ongoing advances in generative AI, enabling clinical informatics teams to harness clinical narrative data effectively without extensive technical expertise or investment in specialized infrastructure.

Computational efficiency and resource optimization

We implemented a dynamic, single-sample allocation strategy that addressed inefficiencies associated with traditional batching methods typically recommended for GPU optimization. In classical machine learning pipelines, batching inputs into fixed-length tensors is considered best practice for computational efficiency.[36] However, in clinical NLP, considerable variability in note lengths causes excessive padding and memory overhead, thus degrading efficiency.[37] By dynamically loading each snippet individually, our pipeline eliminated padding-induced computational bottlenecks, achieving higher processing speeds and more predictable linear scaling of inference time relative to input length. Our approach resulted in significant time

savings when scaled to millions of clinical documents, reducing computational costs and substantially lowering GPU energy consumption and operating temperatures (by 18 ± 2%), thus promoting hardware sustainability and operational longevity.

Novelty and differentiation from existing approaches

SPELL-LLMs significantly differentiated itself from traditional NLP methodologies, including ontology-driven systems such as cTAKES and MetaMap,[28,29] and transformer-based clinical models such as ClinicalBERT or BioBERT.[19,20] While ontology-based systems rely heavily on standardized vocabularies (e.g., UMLS, SNOMED-CT) and manual maintenance, [28,29] they frequently encounter performance limitations when extracting nuanced, rare, or context-dependent clinical concepts.[3] Similarly, transformer-based NER pipelines require substantial domain-specific fine-tuning and annotated datasets, increasing operational complexity and limiting rapid adaptability. [19,20] [18]

A recent study developed a similar regex-LLM framework tailored specifically for spinal surgery data extraction; however, it relied on cloud-based GPT-4 processing, which raises privacy concerns that SPELL-LLMs explicitly address through local hosting.[31] Another recent method introduced clinical entity-based retrieval to enhance extraction efficiency and accuracy, but did not emphasize local model deployment or zero-shot extraction tailored to social determinants of health.[32] Additionally, an evaluation of open-source LLMs demonstrated improved accuracy for extracting social determinants of health compared to traditional regex methods, though without prioritizing local hosting or explicit compliance with healthcare privacy standards.[15] In contrast, SPELL-LLMs uniquely combines regex-based preprocessing with locally hosted generative LLM inference, enabling high computational efficiency, accuracy, zero-shot adaptability, and strict adherence to healthcare privacy regulations.[27]

Potential clinical and operational impact

The substantial (97%) reductions in manual annotation workloads achieved by SPELL-LLMs translated into significant operational improvements, potentially streamlining workflows in clinical research, population health management, and clinical quality assurance. Automating data extraction from clinical narratives not only freed clinicians from resource-intensive annotation tasks but also allowed greater focus on clinical interpretation, intervention planning, and patient care quality improvement initiatives. Furthermore, SPELL-LLMs' demonstrated flexibility and modularity positioned it uniquely for supporting clinical decision-making through real-time or near-real-time extraction of clinical insights from narrative documentation, thereby potentially enhancing patient safety, outcome monitoring, and predictive analytics capabilities.[2,4,5]

Strengths of the proposed approach

A key strength of our approach is the clear separation between regex-based snippet retrieval and LLM-based contextual interpretation. This design allowed domain experts to rapidly refine queries and prompts tailored to new clinical scenarios, significantly enhancing usability and reducing operational barriers. The explicit privacy-focused approach—local hosting and inference—was crucial for addressing data governance challenges faced by healthcare organizations seeking compliant informatics solutions.[13,27] Additionally, our explicit demonstration of generalizability using the MT Samples dataset provided clear evidence of SPELL-LLMs' adaptability beyond the original obstetric scenarios, highlighting the pipeline's potential for immediate adoption across diverse clinical domains. Furthermore, SPELL-LLM has been optimized to efficiently process millions of notes, enabling practical deployment at scale in real-world healthcare settings.

Limitations and future directions

Despite clear feasibility, several limitations remain. Our primary evaluations focused on obstetric-related scenarios, chosen due to their standardized terminologies and structured clinical documentation. Although initial generalizability testing demonstrated promising results for additional conditions (such as ventricular tachycardia), further evaluation across more medical specialties with varied documentation practices is essential for confirming broader applicability. Additionally, our current model (Llama3.1-8B-Instruct), while robust, was not explicitly optimized for clinical language. We prioritized practical clinical deployment, selecting Llama 3.1 after initial evaluations showed superior accuracy, scalability, and practicality compared to medically fine-tuned models (e.g., ClinicalBERT, BioGPT, ClinicalGPT). Although direct comparisons were not extensively pursued here due to known limitations of these specialized models, including shorter context windows and extensive fine-tuning requirements, prior research supports our model selection.[34] Future studies could explore integration with domain-specialized models (e.g., ClinicalBERT variants) to enhance accuracy further.[19,20]

Another limitation was the absence of explicit mapping of extracted textual concepts to standardized clinical terminologies (ICD-10, SNOMED-CT), potentially limiting interoperability. Future enhancements should incorporate explicit concept-mapping modules. Moreover, incorporating explicit linguistic modifier handling (e.g., negation or uncertainty detection via NegEx or FastContext algorithms)[38,39] may further increase extraction reliability.

Further future directions should include external validation across institutions and EHR systems to ensure robustness, proactive monitoring of concept drift to sustain extraction fidelity amidst evolving clinical documentation practices, and demographic fairness analyses to ensure equitable performance across patient populations.

Comparison with existing methodologies

While structured clinical data warehouses and structured query systems (e.g., i2b2) offer systematic extraction alternatives,[33,40] these solutions require significant initial investment and infrastructure adjustments. In contrast, SPELL-LLMs leveraged existing institutional EHR infrastructures and standard NLP libraries, providing immediate practical utility, significantly lower setup costs, and rapid deployment capabilities.

Practical considerations for clinical deployment

Clinician involvement is critical for defining extraction criteria, guiding prompt design, and interpreting ambiguous cases. Simplified guidance or automated parameter tuning (e.g., context window sizing, prompt refinement) could further enhance usability for clinical informatics teams without specialized NLP expertise. SPELL-LLMs' modular Python architecture, structured outputs, and locally hosted design facilitate future interoperability with EHR systems, supporting potential real-time clinical decision-making and improved patient-centered outcomes.

CONCLUSION

SPELL-LLMs demonstrated considerable potential as a robust, efficient, and privacy-compliant NLP pipeline for automating information extraction from EHR narratives. Its modular, scalable design effectively addressed critical informatics challenges, significantly reducing annotation workloads, improving operational efficiency, and enhancing clinical research capabilities. Successful validation within obstetric scenarios demonstrated feasibility, establishing a clear foundation for extending this approach broadly across diverse clinical settings, thus supporting more effective clinical research and enhancing evidence-based decision-making in healthcare.

Acknowledgement: During manuscript preparation the authors used OpenAI ChatGPT (o3 and o4.5 model) to assist with language polishing and consistency checks. The authors reviewed and edited all AI-generated suggestions and are solely responsible for the final content.

Competing Interests: VPK reports funding from the Anesthesia Patient Safety Foundation (APSF), and BWH IGNITE Award. KJG has served as a consultant to Aetion, Roche, BillionToOne, Janssen Global, and Pfizer outside the scope of the submitted work. VPK reports consulting fees from Avania CRO unrelated to the current work. VPK reports patent #WO2021119593A1 for the control of a therapeutic delivery system assigned to Mass General Brigham. DWB reports grants and personal fees from EarlySense, personal fees from CDI Negev, equity from Valera Health, equity from CLEW, equity from MDClone, personal fees and equity from AESOP Technology, personal fees and equity from FeelBetter, and grants from IBM Watson Health, outside the submitted work.

Funding statement: Supported by the NIH/NHLBI 1K08HL161326-01A1 (VPK),NIH/NHLBI R03 HL162756 (KJG) and NIH/NICHD R01 HD117802 (KJG, VPK).

REFERENCES

1. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform. 2017 Sept;73:14–29.

2. Alsentzer E, Rasmussen MJ, Fontoura R, Cull AL, Beaulieu-Jones B, Gray KJ, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. NPJ Digit Med. 2023 Nov 30;6(1):212.
3. Fraile Navarro D, Ijaz K, Rezazadegan D, Rahimi-Ardabili H, Dras M, Coiera E, et al. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. Int J Med Inf. 2023 Sept;177:105122.
4. Liu J, Nguyen A, Capurro D, Verspoor K. Comparing Text-Based Clinical Risk Prediction in Critical Care: A Note-Specific Hierarchical Network and Large Language Models. IEEE J Biomed Health Inform. 2025 May 27;PP.
5. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023 July;619(7969):357–62.
6. Schwarz CM, Hoffmann M, Schwarz P, Kamolz LP, Brunner G, Sendlhofer G. A systematic literature review and narrative synthesis on the risks of medical discharge letters for patients' safety. BMC Health Serv Res. 2019 Mar 12;19(1):158.
7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144–51.
8. Yim W wai, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol. 2016 June 1;2(6):797.
9. Neves M, Leser U. A survey on annotation tools for the biomedical literature. Brief Bioinform. 2014 Mar 1;15(2):327–40.
10. Irrera O, Marchesin S, Silvello G. MetaTron: advancing biomedical annotation empowering relation annotation and collaboration. BMC Bioinformatics. 2024 Mar 14;25(1):112.
11. Neves M, Ševa J. An extensive review of tools for manual annotation of documents. Brief Bioinform. 2021 Jan 18;22(1):146–63.

12. Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. J Biomed Inform. 2015 Apr;54:186–90.
13. Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. Npj Digit Med. 2024 Aug;7(1):222.
14. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication. JAMA Surg. 2024 Aug;159(8):928.
15. Gu B, Shao V, Liao Z, Carducci V, Brufau SR, Yang J, et al. Scalable information extraction from free text electronic health records using large language models. BMC Med Res Methodol. 2025 Jan;25(1):23.
16. Samhammer D, Roller R, Hummel P, Osmanodja B, Burchardt A, Mayrdorfer M, et al. "Nothing works without the doctor:" Physicians' perception of clinical decision-making and artificial intelligence. Front Med. 2022 Dec;9:1016366.
17. Simoulin A, Thiebaut N, Neuberger K, Ibnouhsein I, Brunel N, Viné R, et al. From free-text electronic health records to structured cohorts: Onconum, an innovative methodology for real-world data mining in breast cancer. Comput Methods Programs Biomed. 2023 Oct;240:107693.
18. Tavabi N, Pruneski J, Golchin S, Singh M, Sanborn R, Heyworth B, et al. Building large-scale registries from unstructured clinical notes using a low-resource natural language processing pipeline. Artif Intell Med. 2024 May;151:102847.
19. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019 [cited 2025 Mar 26]. p. 72–8. Available from: http://aclweb.org/anthology/W19-1909

20. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinforma Oxf Engl. 2020 Feb 15;36(4):1234–40.
21. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North [Internet]. Minneapolis, Minnesota: Association for Computational Linguistics; 2019 [cited 2025 Mar 26]. p. 4171–86. Available from: http://aclweb.org/anthology/N19-1423
22. Sciannameo V, Pagliari DJ, Urru S, Grimaldi P, Ocagli H, Ahsani-Nasab S, et al. Information extraction from medical case reports using OpenAI InstructGPT. Comput Methods Programs Biomed. 2024 Oct;255:108326.
23. Bhagat N, Mackey O, Wilcox A. Large Language Models for Efficient Medical Information Extraction. AMIA Jt Summits Transl Sci Proc. 2024;2024:509–14.
24. Gu Z, Jia W, Piccardi M, Yu P. Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought. Artif Intell Med. 2025 Apr;162:103078.
25. Ai Q, Bai T, Cao Z, Chang Y, Chen J, Chen Z, et al. Information Retrieval meets Large Language Models: A strategic report from Chinese IR community. AI Open. 2023;4:80–90.
26. Yang Z, Xu SS, Liu X, Xu N, Chen Y, Wang S, et al. Large Language Model–Based Critical Care Big Data Deployment and Extraction: Descriptive Analysis. JMIR Med Inf. 2025 Mar;13:e63216–e63216.
27. Wiest IC, Ferber D, Zhu J, Van Treeck M, Meyer SK, Juglan R, et al. Privacy-preserving large language models for structured medical information retrieval. Npj Digit Med. 2024 Sept;7(1):257.
28. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17–21.

29. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc JAMIA. 2010;17(5):507–13.
30. Abbas A, Afzal M, Hussain J, Ali T, Bilal HSM, Lee S, et al. Clinical Concept Extraction with Lexical Semantics to Support Automatic Annotation. Int J Environ Res Public Health. 2021 Oct 9;18(20):10564.
31. Dagli MM, Ghenbot Y, Ahmad HS, Chauhan D, Turlip R, Wang P, et al. Development and validation of a novel AI framework using NLP with LLM integration for relevant clinical data extraction through automated chart review. Sci Rep. 2024 Nov 5;14(1):26783.
32. Lopez I, Swaminathan A, Vedula K, Narayanan S, Nateghi Haredasht F, Ma SP, et al. Clinical entity augmented retrieval for clinical information extraction. NPJ Digit Med. 2025 Jan 19;8(1):45.
33. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc AMIA Symp. 2006;2006:1040.
34. Rohanian O, Nouriborji M, Kouchaki S, Nooralahzadeh F, Clifton L, Clifton DA. Exploring the effectiveness of instruction tuning in biomedical language processing. Artif Intell Med. 2024 Dec;158:103007.
35. MT Samples Dataset [Internet]. [cited 2025 May 1]. Available from: www.mtsamples.com
36. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc; 2019.
37. Liu J, Capurro D, Nguyen A, Verspoor K. Uncovering Variations in Clinical Notes for NLP Modeling. Stud Health Technol Inform. 2024 Jan 25;310:1460–1.
38. Shi J, Hurdle JF. Trie-based rule processing for clinical NLP: A use-case study of n-trie, making the ConText algorithm more efficient and scalable. J Biomed Inform. 2018 Sept;85:106–13.

39. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, et al. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform. 2015 Apr;54:213–9.

40. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc JAMIA. 2010 Apr;17(2):124–30.

FIGURES

Figure 1. Overview of the SPELL-LLM platform. The system comprises four key components: (1) Data request—clinicians and researchers retrieve clinical data from institutional database repositories; (2) Metadata extraction—a preprocessing step that indexes reports and extracts key metadata to streamline cohort selection; (3) Regex-based search—a filtering module that eliminates irrelevant reports and extracts snippets from relevant matches; and (4) LLM-based retrieval—a locally hosted large language model that analyzes extracted snippets to enhance information retrieval efficiency. This iterative workflow enables precise alignment between selected patient cohorts and research objectives.

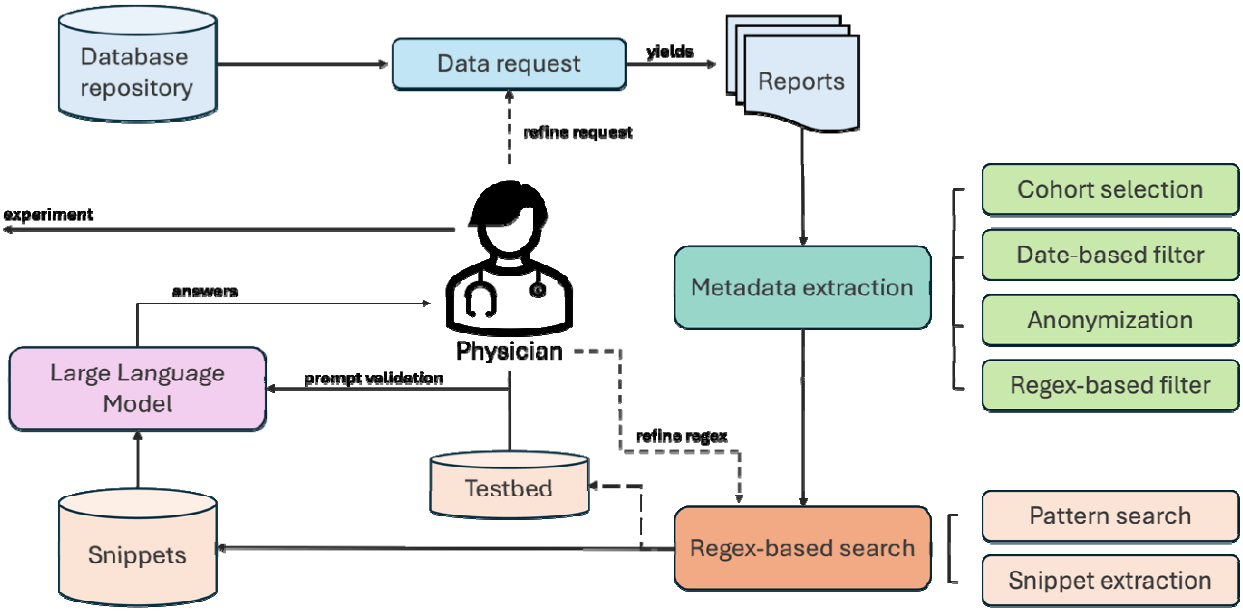

Figure 2. Overview of the system workflow in SPELL-LLM, illustrating the functionality of each module through a representative example present in the EDD extraction task. Each block demonstrates a specific module in action, with arrows indicating the sequential processing steps from dataset creation to information extraction.

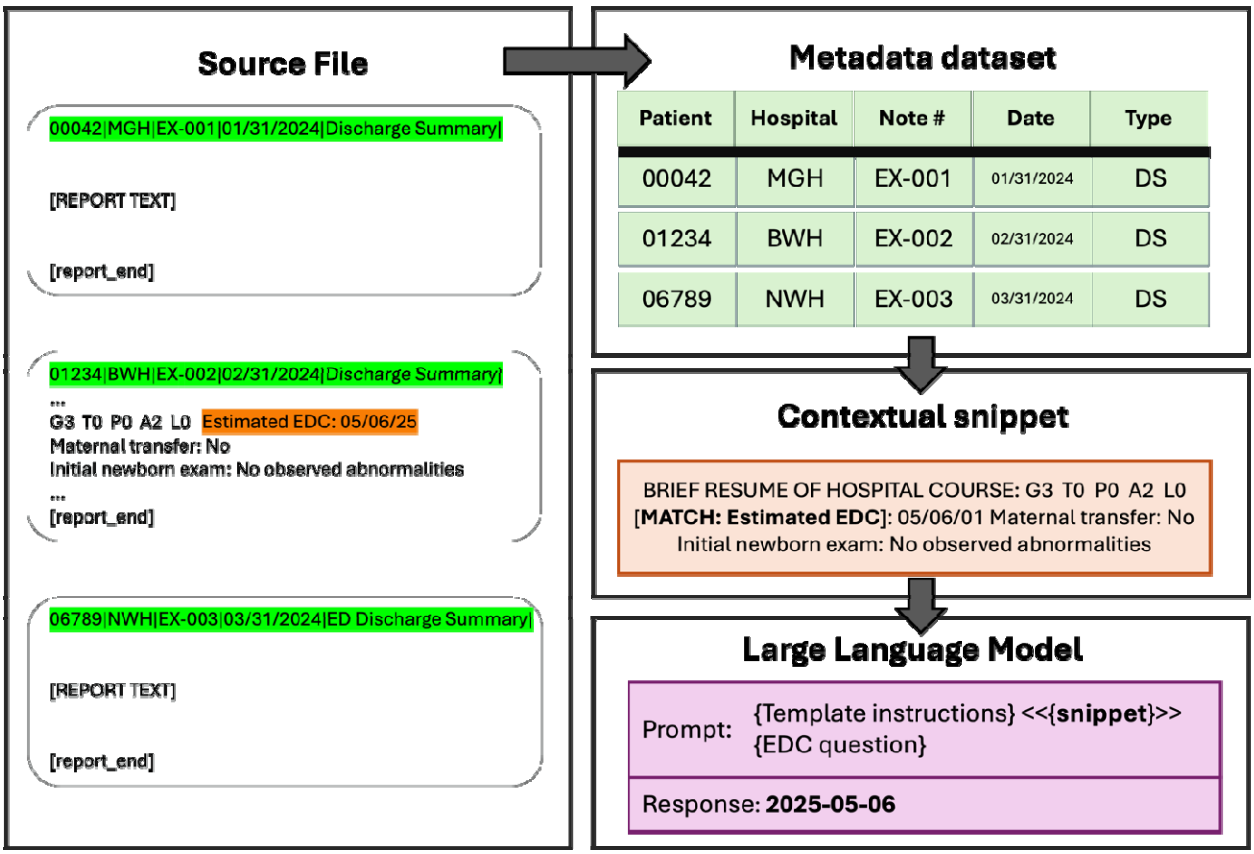
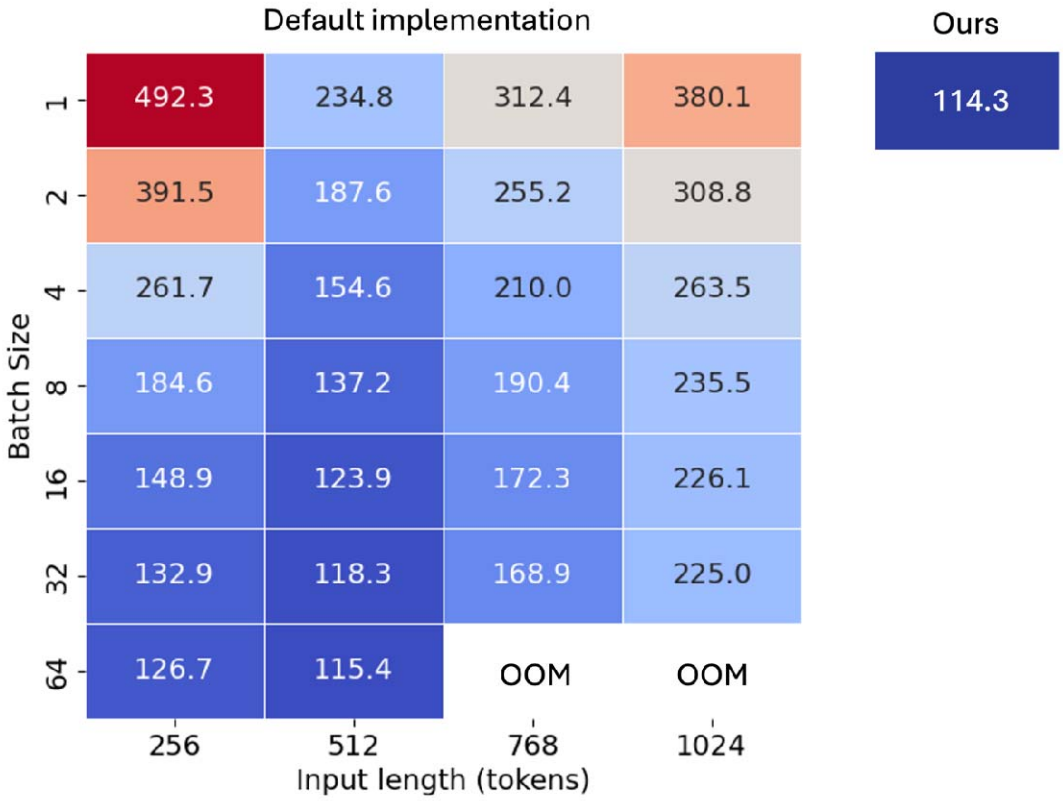
Figure 3: Total processing time (in seconds) for 540 clinical notes across different batch sizes (y-axis) and input lengths in tokens (x-axis) using the preferred implementation of Hugging Face pipeline's API. Each cell represents the total runtime for a specific combination of these hyperparameters. Two configurations resulted in Out-Of-Memory (OOM) errors. Shown on the right, a standalone cell displays the total processing time achieved using our custom pipeline, which dynamically handles single-sample input length to optimize efficiency.